Dynamic Classification for Administrative Health Data Case Definitions: Application to Juvenile Arthritis

by

Allison Feely

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Community Health Sciences

University of Manitoba

Winnipeg

Copyright © 2019 by Allison Feely

ABSTRACT

Introduction: Administrative health databases are important resources for population-based chronic disease research and surveillance, but case definitions to identify disease cases can have low sensitivity and specificity. Case definitions constructed using repeated measurements of diagnoses over time have potential for improved accuracy. In particular, dynamic classification, which uses probabilistic models to classify individuals into disease, non-disease, and indeterminate categories and then updates these classifications as new information becomes available, are promising alternatives to conventional static classification methods for case ascertainment.

Purpose & Objectives: The research purpose was to develop and evaluate methods that use longitudinal diagnostic information from population-based administrative databases to improve the accuracy of chronic disease case definitions. The objectives were to: (1) develop dynamic classification methods for retrospective longitudinal administrative data, (2) apply and validate dynamic classification to identify cases of juvenile arthritis (JA), and (3) compare the performance of case definitions developed using dynamic classification with case definitions developed using static approaches for classification.

Methods: Dynamic longitudinal discriminant analysis (LoDA) was adapted for retrospective longitudinal administrative data and applied to administrative health data from Manitoba to identify cases of JA. The Pediatric Rheumatology Clinical Database was used for validation. Performance of the JA case definitions constructed with dynamic LoDA was measured using sensitivity, specificity, positive predictive value (PPV), and mean time to classification. Classification accuracy was compared for dynamic LoDA, deterministic case definitions and static LoDA models.

Results: The study cohort included 797 children from the clinical database who could be linked to administrative health data from birth to age 16; 386 (48.4%) were JA cases and 411 (51.6%) were non-cases. The dynamic LoDA model with the best fit used a longitudinal binary variable for any-JA related physician visit or hospitalization. It had sensitivity of 0.70, specificity of 0.81, PPV of 0.82, and left 2% of the cohort unclassified after all data were used. On average, it took 9.21 years of data to classify individuals as a JA case or non-case. Both the deterministic case definition and static LoDA model outperformed the best dynamic LoDA model.

Conclusion: The results suggest that dynamic classification can produce accurate case definitions using longitudinal information from administrative health data, although the choice of methods and their comparative performance will depend on the characteristics of the disease.

ACKNOWLEDGEMENTS

First, I would like to sincerely thank my supervisor, Dr. Lisa Lix, for her guidance and support throughout my Master's program. She is a true mentor and I will always be grateful for everything she has taught and provided me. I would also like to thank my committee members, Dr. Lily Lim and Dr. Depeng Jiang, for providing constructive feedback throughout the development of my thesis. My thesis would be not be complete without their clinical and methodological expertise.

I greatly appreciate the workspace and resources that were provided to me by the George and Fay Yee Centre for Healthcare Innovation. It was wonderful to work alongside so many talented and kind individuals. I am also grateful for the financial support I received from the Manitoba Graduate Scholarship, the Visual and Automated Disease Analytics (VADA) Training Program Fellowship, and the Canadian Institutes of Health Research (CIHR) Canada Graduate Scholarship. In addition, I would like to thank Dr. Leslie Roos for providing me with the opportunity to work on his research, which allowed me to gain invaluable skills using administrative data.

Finally, I would like to thank my family and friends. None of this would have been possible without their support. A special thanks to my mom and dad for their constant encouragement throughout my entire education and letting me live at home "rent-free".

The results and conclusions are those of the authors and no official endorsement by the Manitoba Centre for Health Policy, Manitoba Health, Seniors and Active Living, or other data providers is intended or should be inferred.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Definition
AUC	Area under the curve
CCDSS	Canadian Chronic Disease Surveillance System
CrI	Credible interval
DAD	Discharge Abstract Database
EHR	Electronic health record
FNR	False negative rate
FPR	False positive rate
GRU	Gated recurrent unit
HPD	Highest posterior density
IC/ES	Institute for Clinical Evaluative Sciences
ICD-10-CA	International Classification of Diseases, Tenth Revision, Canada
ICD-9	International Classification of Diseases, Ninth Revision
ICD-9-CM	International Classification of Diseases, Ninth Revision, Clinical Modification
ILAR	International League of Associations for Rheumatology
JA	Juvenile arthritis
JIA	Juvenile idiopathic arthritis
JRA	Juvenile rheumatoid arthritis
LoDA	Longitudinal discriminant analysis
LSTM	Long short-term memory
MCHP	Manitoba Centre for Health Policy
MCMC	Markov chain Monte Carlo
MGLMM	Multivariate generalized linear mixed model
NPV	Negative predictive value
PCC	Probability of correct classification
PED	Penalized expected deviance
PPV	Positive predictive value
PSRF	Potential scale reduction factor
RNN	Recurrent neural network
ROC	Receiver operating characteristic curve
SANAD	Standard and New Antiepileptic Drug study
SEA	Seronegative enthesopathy and arthropathy syndrome

CHAPTER 1 – INTRODUCTION

1.1 Background

Administrative health databases capture electronic information about contacts with the health care system, such as physician visits, hospitalizations, and prescription drug dispensations. These data were originally collected for purposes of managing and monitoring the health care system, but are increasingly used for secondary purposes, including chronic disease research and surveillance. Population-based studies using administrative health databases have been conducted for a wide variety of chronic diseases, including hypertension, osteoporosis, and inflammatory bowel disease (Bernstein et al., 2006; O'Donnell & Canadian Chronic Disease Surveillance System (CCDSS) Osteoporosis Working Group, 2013; Robitaille et al., 2012).

Administrative health databases are useful sources of data for chronic disease research and surveillance because they are relatively inexpensive to access, contain standardized diagnosis and treatment information, can be linked to create longitudinal records of health care use for individuals, and capture routinely collected information on entire populations (Cadarette & Wong, 2015; Jutte et al., 2011; Virnig & McBean, 2001). However, the quality of administrative data is an ongoing concern (Smith et al., 2017). This includes measurement error (i.e., misclassification error) in diagnosis codes, which can lead to biased estimates of chronic disease incidence and prevalence (Manuel et al., 2010; Wilchesky et al., 2004).

Individuals with chronic diseases are identified in administrative health databases by case definitions, the rules used to assign an individual as a disease case or non-case. Ideally, case definitions are validated using clinically confirmed case information. Deterministic and model-based approaches can be used to develop case definitions for chronic diseases. The deterministic approach involves *a priori* development of the rules used to identify disease cases from administrative databases. These rules include the number and types of diagnostic/medication codes that need to be present in administrative records in a specified period of time for an individual to be identified as a disease case (Lix et al., 2006). For example, a commonly used deterministic case definition for hypertension is “1 or more hospitalization or 2 or more physician claims within a 2 year-period” with relevant hypertension diagnosis codes from the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) and International Classification of Diseases, Tenth Revision, Canada (ICD-10-CA) (ICD-9-CM codes: 401–405; ICD-10-CA codes: I10–I13, I15) (Quan et al., 2009). The deterministic approach is popular since it is easy to apply; however, case definitions based solely on diagnosis codes are sensitive to misclassification bias in these codes (Lix et al., 2008; Manuel et al., 2010; Van Walraven et al., 2010).

The model-based approach uses statistical or machine learning models to identify disease cases from administrative health data. Case definitions developed using the model-based approach can include related diagnosis/medication information, but can also incorporate additional information recorded in administrative data that may be important for identifying disease cases, such as age, sex, and diagnoses for comorbid conditions. It has been shown that models that include multiple features from administrative health data can improve the accuracy of case ascertainment (Cooke et al., 2011; English et al., 2016; Fan et al., 2013; Lix et al., 2008; Peng et al., 2015; Van Walraven et al., 2010).

However, there has been little, if any, research regarding the use of repeated measurements of information over time in administrative health data to develop case definitions. Currently, most case definitions that incorporate longitudinal information do so using an aggregated approach, by counting the number of diagnosis/medication codes that are recorded in administrative databases in a specified period of time. Classification models that leverage the temporal relationships in longitudinal data have increased prediction accuracy when compared with aggregated models based on the area under the curve (AUC) of the receiving operator characteristic (ROC) (Choi et al., 2017; Hughes, Komárek, et al., 2018; Maruyama et al., 2009; Reddy & Delen, 2018; Wang et al., 2018). One potentially useful model-based method for developing chronic disease case definitions that uses repeated measurements is dynamic classification (Hughes et al., 2017; Hughes, Komárek, et al., 2018). Dynamic classification relies on probabilistic models to classify individuals into disease, non-disease, and indeterminate categories and then updates the classifications of the individuals in the indeterminate category as new information becomes available.

Dynamic classification methods were developed for prospective longitudinal clinical studies with the objective of potentially identifying some cases earlier, which could improve patient care by allowing for earlier treatment (Hughes et al., 2017; Hughes, Komárek, et al., 2018). However, to date, these methods have not been extended to the retrospective setting, specifically to retrospective cohort studies using longitudinal administrative health data. Differences in how dynamic classification methods are applied to retrospective studies compared to prospective studies may exist. For example, in prospective studies, classifications are updated each time new information is collected. However, in retrospective studies using administrative health data, all of the longitudinal data has been collected before the study begins. Thus, a topic for exploration is when the classification should be updated. One approach is to update classifications each time a new visit is recorded in the administrative health data, similar to the approach used in the prospective setting (Hughes et al., 2017; Hughes, Komárek, et al., 2018). This method uses a person-specific approach to updating classifications. A second approach is to update the classifications at fixed points in time, such as monthly or annually. This approach requires consideration of the optimal time to update, which could depend on the outcome of interest or the characteristics of the data.

1.2 Purpose and Objectives

The purpose of this research was to develop and evaluate methods that use longitudinal information (i.e., repeated measurements) from population-based administrative databases to improve the accuracy of chronic disease case definitions. The objectives were:

1. To develop dynamic classification methods for retrospective longitudinal administrative health data.
2. To apply and validate dynamic classification methods in a real-world example using administrative health data to identify cases of juvenile arthritis (JA).
3. To compare the performance of JA case definitions developed using dynamic classification with JA case definitions developed using a deterministic approach and a static model-based approach.

1.3 Thesis Organization

Chapter 2 includes a literature review on the following topics: (1) model-based methods for developing case definitions using cross-sectional and longitudinal data, (2) classification structures that include an indeterminate category, and (3) studies that developed JA case definitions for administrative health data. Chapter 3 describes the methods for dynamic classification using longitudinal administrative health data, as well as for the JA application. Chapter 4 presents the results from the JA application, and the last chapter discusses key findings, conclusions, and opportunities for future research.

CHAPTER 2 – LITERATURE REVIEW

This literature review includes the following topics: model-based methods for developing case definitions using cross-sectional and longitudinal data, classification structures that include an indeterminate category, and studies that developed JA case definitions for administrative health data.

2.1 Model-Based Methods for Developing Case Definitions

2.1.1 Cross-Sectional Data

Various model-based approaches have been applied to administrative data for case ascertainment. The majority of studies on this topic used logistic regression to classify individuals as disease cases and non-cases based on diagnosis/procedure codes and other relevant measures recorded in administrative databases, such as age, sex, and presence of comorbid conditions (Cooke et al., 2011; English et al., 2016; Lix et al., 2008; Peng et al., 2015; Van Walraven et al., 2010). In addition to logistic regression, Lix et al. (2008) applied artificial neural networks and classification trees to administrative data to identify cases of osteoporosis in Manitoba women aged 50 years and older, and English et al. (2016) used recursive partitioning to identify hospitalizations for primary subarachnoid hemorrhage. Artificial neural networks are machine learning models that are constructed based on the function of the brain. Classification trees or recursive partitioning are nonparametric methods for classification where the data are split into homogeneous sub-groups (i.e., leaves) based on several predictor variables. Each of the leaves are then assigned to one of several classification groups of interest. Each of these studies aggregated the diagnosis and procedure information for the entire study period by either creating binary variables for the presence or absence of specific diagnosis/procedure codes or by counting the number of diagnosis/procedure codes that were present in administrative records.

2.1.2 Longitudinal Data

Several studies have proposed methods for model-based case ascertainment that leverage the temporal information in repeated measures data. These longitudinal approaches aim to model the change in an individual's characteristics (e.g., risk factors and healthcare system utilization) over time to predict health outcomes.

One method for model-based case ascertainment using repeated measures data is longitudinal discriminant analysis (LoDA). Discriminant analysis is a multivariate statistical technique for classifying individuals into groups. Traditionally, discriminant analysis is applied to multivariate data for a single point in time or time period and includes only baseline and cross-sectional characteristics. However, it can also be applied to repeated measures data (Lix & Sajobi, 2010). Most methods of LoDA use mixed effects models applied to the repeated measures data, as they are able to account for the correlation between successive measurements. The flexibility of mixed effects models to include both time-varying and time-invariant covariates, as well as their ability to accommodate unequal and asynchronous measurements make them useful for clinical applications (Hughes, Komárek, et al., 2018; Lix & Sajobi, 2010).

Several studies used LoDA with linear mixed effects models for a single continuous longitudinal variable (Lix & Sajobi, 2010). The LoDA methodology was expanded to allow multiple continuous variables to be simultaneously modelled using multivariate mixed effects models. Marshall et al. (2009) used a non-linear mixed effects model with two longitudinal continuous variables to predict normal and abnormal pregnancy outcomes. Komárek et al. (2010) used a multivariate linear mixed effects model to jointly model three continuous longitudinal variables to predict the prognosis of primary biliary cirrhosis patients. A normal mixture for the random effects distribution was introduced in the mixed effects model to relax the traditional normality assumption. Morrell et al. (2012) used LoDA with a multivariate mixed effects model to predict cases of prostate cancer using three continuous longitudinal variables. Hughes et al. (2018) extended the multivariate LoDA approach to allow for longitudinal variables of different types (i.e., continuous, binary, count) by adopting a multivariate generalized linear mixed model (MGLMM). This MGLMM approach has been used to identify patients who will not achieve remission of seizures (Hughes, Bonnett, et al., 2018; Hughes et al., 2017; Hughes, Komárek, et al., 2018) and to identify diabetic patients who will develop sight-threatening diabetic retinopathy (García-Fiñana et al., 2018).

A second case ascertainment method for repeated measures data is the shared random effects model. These models consist of two submodels: (1) a linear mixed effects submodel for the trajectories of a continuous predictor for individuals over time, and (2) a generalized linear submodel for the primary outcome (e.g., binary event). These two submodels are linked by shared random effects coefficients. Two model estimation approaches were developed. The first is a two-stage approach where the submodels are estimated separately (Albert, 2012; Maruyama et al., 2009; Zeng et al., 2014). First, the linear mixed effects model is fit to the longitudinal data and estimates are produced for the random slopes and intercepts. Second, in the case of a binary outcome, a logistic regression model is fit using the estimated random slopes and intercepts as covariates. The second estimation method jointly models the two submodels using a Bayesian approach where the parameters of the two submodels are estimated using Markov Chain Monte Carlo (MCMC) methods (Elliott et al., 2012; Horrocks & van Den Heuvel, 2009; Jiang et al., 2015; Mohd Din et al., 2014). The two-stage modelling approach can be easily implemented using standard regression software, such as SAS. However, the two-stage approach does not account for the error in the estimation of the random effects coefficients and is sensitive to measurement error. This measurement error is corrected for in the joint modelling approach; however, implementation of the joint approach is more computationally intensive (Horrocks & van Den Heuvel, 2009; Mohd Din et al., 2014).

Case ascertainment using repeated measures data can also be conducted using machine learning models. Traditional machine learning methods are not designed to extract the temporal relationships that exist in longitudinal data, but recurrent neural networks (RNNs) include a feedback loop that allow information from previous time points to be retained. This makes them an effective method to model repeated measures data (Xiao et al., 2018; Yao et al., 2018). Two variations of RNNs are prominently used to predict clinical outcomes using longitudinal electronic health record (EHR) data: gated recurrent unit (GRU) and long short-term memory (LSTM) (Xiao et al., 2018; Yao et al., 2018).

Recently, Choi et al. (2017) used a GRU to predict the initial diagnosis of heart failure, Reddy and Delen (2018) used a LSTM to predict rehospitalization within 30 days for Lupus patients, Wang et al. (2018) used a LSTM to predict the progression stage of Alzheimer's disease, and Duan et al. (2019) used a bi-directional LSTM to predict main adverse cardiac events for acute coronary syndrome patients during hospitalization. All of these studies determined that the RNN models performed better than the traditional approaches that ignore temporality, such as logistic regression and support vector machines. Although RNN models are a useful approach for case ascertainment using longitudinal information, they do present some limitations, such as the inability to handle missing data and challenges in interpreting the importance of the predictors (Reddy & Delen, 2018).

2.2 Classification Structure

Traditional statistical classification involves assigning participants to one of two or more categories based on data features. Since classification decisions are based on probabilistic models, they are subject to error and participants may be misclassified. Misclassifications may result because there is insufficient information about an individual. Introducing an indeterminate category into the classifier may increase accuracy. An individual could be assigned to the indeterminate category rather than one of the defined categories if they lack the adequate information to be confidently classified. Instead, the individual is referred to follow-up where additional information is used to make a potentially more accurate classification. Current classification methodologies that adopt the indeterminate category are neutral zone classifiers (Hua et al., 2010; Jeske et al., 2017, 2007; Jeske & Smith, 2018; Zhang et al., 2016) and dynamic classification (Hughes et al., 2017; Hughes, Komárek, et al., 2018).

2.2.1 Neutral Zone Classifiers

Neutral zone classifiers add a “neutral” (i.e., indeterminate) classification group to screen out individuals that are difficult to classify with existing information. A neutral zone can make a classifier more useful because it controls the false positive rate (FPR) and false negative rate (FNR). Neutral zone classifiers were motivated by applications in microbial community profiling (Hua et al., 2010; Jeske et al., 2007) and were more recently applied to predict prostate cancer reoccurrence (Jeske et al., 2017; Jeske & Smith, 2018) and kidney dysfunction following heart surgery (Jeske & Smith, 2018; Zhang et al., 2016).

Jeske et al. (2007) defined the two-class neutral zone framework as:

$$d_B(y;\ L_0, L_1) = \begin{cases} 0, & \text{if } P(C = 1|Y = y) \leq L_0 \\ N, & \text{if } L_0 < P(C = 1|Y = y) < L_1 \\ 1, & \text{if } P(C = 1|Y = y) \geq L_1 \end{cases} \tag{2.1}$$

where $P(C = 1|Y = y)$ is the posterior probability of the event $C = 1$ given the data and $(L_0, L_1) \in D$ with $D = \{(L_0, L_1): 0 \leq L_0 \leq L_1 \leq 1\}$. Hua et al. (2010) extended the neutral zone framework to the three-class classification problem. Equation (2.1) shows that an individual is classified into the neutral (*N*) category if their posterior probability is neither large enough to be classified into group 1 nor small enough to be classified into group 0, meaning they have weak evidence supporting their classification. The values for L_0 and L_1 are chosen by quantifying the amount of certainty that is needed for the user to feel confident in making classification decisions. This is done by minimizing a cost structure described by the FPR and FNR. Jeske and Smith (2018) also developed a method for building a neutral zone classifier using a Receiving Operator Characteristic (ROC) curve, which allows neutral zone regions to be used with logistic regression.

2.2.2 Dynamic Classification

Dynamic classification methods include an “unclassified” (i.e., indeterminate) category for individuals with insufficient evidence for classification. The motivation behind dynamic classification is that some individuals may have sufficient evidence to be classified into a group at an earlier time point than others, which would potentially allow for earlier treatment (Hughes et al., 2017; Hughes, Komárek, et al., 2018).

Hughes, Komárek, et al. (2018) developed dynamic LoDA to use an individual’s longitudinal clinical history to accurately classify individuals into two groups (e.g., Group 0 and Group 1) as early as possible. The LoDA is based on multivariate generalized linear mixed models (MGLMMs) to allow multiple types of longitudinal variables (e.g., continuous, binary, count) to be jointly modelled. The dynamic LoDA procedure involves updating an individual’s group membership probability at each time point that new data are available and is adapted from the sequential classification approach previously used to predict the development of prostate cancer (Brant et al., 2003; Morrell et al., 2012) and Alzheimer's disease (Brant et al., 2005). At a given time point, if an individual's probability of belonging to Group 1 is greater than a cut-off, c, that individual is classified into Group 1 and the classifications stop for that individual. Otherwise, the individual is considered "unclassified" and the classification procedure is repeated at the next time point. This procedure continues until the final time point, where new data are available. The cut-off, c, is chosen by the investigator using an appropriate method that has been specified *a priori* (e.g., point nearest to the top left corner of the ROC curve). Significant improvements in predictive accuracy were obtained with the dynamic LoDA procedure compared to traditional discriminant analysis models (Hughes, Komárek, et al., 2018). Without sacrificing predictive accuracy, dynamic LoDA was also able to classify individuals significantly earlier than waiting until the last time point when all the data were collected.

Hughes et al. (2017) expanded the dynamic LoDA methodology by introducing an allocation scheme using the credible intervals (CrIs) of the group membership probabilities. The CrI scheme was developed to improve positive predictive value (PPV) of the dynamic LoDA by reducing the number of false positive and false negatives that occur when making classification decisions based on point estimates. The additional information provided by the CrIs allowed the classification scheme to account for the potential heterogeneity in the precision of the group membership probabilities between individuals. Individuals are only classified if the cut-off, c, falls outside of the $(1-\alpha)100\%$ CrI of their group membership probability. If c falls within the $(1-\alpha)100\%$ CrI, the individual is considered "unclassified" and is referred to the next time point where further information is available. The dynamic CrI method improved PPV without sacrificing sensitivity, specificity, and probability of correct classification (PCC) compared to dynamic LoDA. However, compared to dynamic LoDA, there was a slight delay in the classification time, as the CrI allocation scheme is more conservative.

Dynamic classification was applied to longitudinal data from the Standard and New Antiepileptic Drug (SANAD) study to identify patients with epilepsy unable to achieve a 12-month seizure remission within 5 years of starting treatment (Hughes, Bonnett, et al., 2018). MGLMMs were fit separately for the remission and no-remission groups and were used to calculate the probability of a patient not achieving remission at each time point. Four longitudinal variables were considered, along with multiple potential covariates to model the change in the longitudinal variables over time. Covariates were selected using penalized expected deviance with a forward selection approach and the optimal combination of longitudinal variables was selected based on the PCC. The covariates included in the multivariate models differed between the remission and no-remission groups. The CrI allocation scheme was used by calculating 99% CrIs for the probabilities of not achieving remission. The final dynamic classification model left 5% of the sample unclassified. With these unclassified patients removed, the final model had a sensitivity, specificity, and PPV of 95%, 97%, and 76%, respectively.

2.3 Validated Juvenile Arthritis Case Definitions

Three studies validated JA case definitions using administrative health data; two using data from Canada (Shiff et al., 2017; Stringer & Bernatsky, 2015) and one using data from the United States (Harrold et al., 2013). These studies developed deterministic case definitions to identify cases of JA in physician billing claims, hospital discharge records, and pharmacy data. Diagnoses were recorded in the administrative health databases using ICD-9, ICD-9-CM, and ICD-10-CA codes. Chart review (Harrold et al., 2013) and provincial pediatric rheumatology clinical databases (Shiff et al., 2017; Stringer & Bernatsky, 2015) were used as reference standards.

Harrold et al. (2013) tested seven JA case definitions using administrative health data from California. Cases of JA were identified with ICD-9 codes 696.0, 714, and 720. The estimates of sensitivity and PPV ranged from 81% to 95% and 45% to 91%, respectively. Stringer and Bernatsky (2015) tested four JA case definitions using physician billing claims from Nova Scotia and identified JA cases with ICD-9 code 714; sensitivity estimates ranged from 53% to 86% and PPV estimates ranged from 52% to 65%.

Shiff et al. (2017) validated eight JA case definitions using administrative health data from Manitoba with ICD-9-CM codes 714 and 720 and ICD-10-CA codes M05, M06, M08 and M45. The case definitions were applied to two validation cohorts: one where individuals diagnosed with seronegative enthesopathy and arthropathy (SEA) syndrome were classified as cases of JA and the other where individuals diagnosed with SEA syndrome were classified as non-cases. No significant difference in the accuracy of the case definitions were found between the cohorts. Estimates of sensitivity, specificity, and PPV ranged from 82% to 92%, 81% to 93%, and 88% to 95%, respectively.

2.4 Summary of Literature Review

Model-based approaches with cross-sectional data have been used to develop case definitions for administrative health data by aggregating diagnosis and procedure information over the study periods. Longitudinal model-based classification approaches that model the change in an individual’s characteristics can also be used to develop case definitions. These classification methods leverage the temporal information in repeated measures data and can have increased predictive accuracy compared to models that only include aggregated information.

In addition, classification methods that include an indeterminate category have the potential to increase classification accuracy by not classifying individuals prematurely. Dynamic classification assigns individuals to an "unclassified" category if there is insufficient evidence to support their classification and then refers them to follow-up where additional information can be used to update their classification.

Three studies previously developed and validated JA case definitions for administrative health data. All three of the studies used a deterministic approach to develop the case definitions. The accuracy of the case definitions in identifying JA cases was evaluated using sensitivity, specificity, and PPV.

Based on this review of the literature, there is an opportunity to consider another approach to case ascertainment for JA. This approach relies on dynamic classification of disease markers in longitudinal administrative health data.

CHAPTER 3 – METHODS

This research was conducted by applying dynamic classification methods to administrative health data linked to a clinical database to identify cases of JA in Manitoba. First, this chapter describes the data sources, study cohort, and study variables. Second, a description of the dynamic classification methods that were applied to the retrospective longitudinal administrative health data is provided. Finally, the statistical analysis for the JA application is described.

3.1 Data Sources

This study was conducted using linked administrative health databases and clinical registry data from Manitoba for fiscal years 1980/81 to 2017/18 (a fiscal year extends from April 1 to March 31). The administrative health databases included the Manitoba Health Insurance Registry, Hospital Discharge Abstract Database (DAD), and Medical Claims database. Pediatric rheumatology clinical registry data was used for validation. These databases are contained in the Manitoba Population Research Data Repository housed at the Manitoba Centre for Health Policy (MCHP) and can be linked using a unique anonymized personal health identification number. In addition, Statistics Canada Census data were used, at a dissemination area (i.e., aggregate) level, to define income quintiles. Previous research has used the data in the Repository to construct case definitions for a variety of chronic conditions, including diabetes, hypertension, osteoporosis, and rheumatoid arthritis (Dart et al., 2011; Leslie et al., 2011; Lix et al., 2006; Shiff et al., 2017).

The Manitoba Health Insurance Registry contains individual-level information for residents of Manitoba registered with Manitoba Health at any point since 1970. The registry captures dates of health insurance coverage in Manitoba, reasons for termination of coverage (e.g., migration out of province and death), and demographic characteristics (e.g., age, sex and location of residence).

The DAD contains clinical information, including diagnosis and procedure codes for patients at the point of discharge from Manitoba hospitals. Up to 16 ICD-9-CM diagnosis codes are listed in the DAD until March 31, 2004. From April 1, 2004 onward, up to 25 ICD-10-CA diagnosis codes are listed in the DAD.

The Medical Claims database contains information about physician billings for in-hospital and outpatient visits. Each claim contains a single ICD-9-CM diagnosis code.

The Pediatric Rheumatology Clinical Database is the clinical registry used for validation (Shiff et al., 2017). It contains a confirmed clinical diagnosis for children who were seen by a pediatric rheumatologist at the Children's Hospital in Winnipeg, Manitoba. This clinical registry contains information for each patient seen between 1980 and 2012, including their diagnosis and diagnosis date. Diagnoses are assigned based on information about patient history, physical examination, blood work, imaging, and other investigations performed.

3.2 Study Cohort

The cohort used to develop and validate the dynamic classification case definition for JA was created from the Pediatric Rheumatology Clinical Database. JA is defined generally as arthritis that begins before the 16th birthday and persists for at least six weeks (Petty et al., 2004). Currently, JA is called juvenile idiopathic arthritis (JIA) and is defined by the International League of Associations for Rheumatology (ILAR) classification system (Petty et al., 2004). JIA includes seven subtypes: systemic arthritis, oligoarthritis, polyarthritis rheumatoid factor negative, polyarthritis rheumatoid factor positive, psoriatic arthritis, enthesitis-related arthritis, and undifferentiated arthritis. Prior to the ILAR classification, JA was defined in North America using the 1977 American College of Rheumatology classification for juvenile rheumatoid arthritis (JRA), which included three subtypes: oligoarticular, polyarticular, and systemic (Brewer et al., 1977). In this study, JA was defined using both the JIA and JRA classifications.

The study observation period extended from April 1, 1980 to March 31, 2018. The study cohort included Manitoba residents who met the following inclusion criteria: (1) were born between April 1, 1980 and March 31, 2002; (2) were recorded in the Pediatric Rheumatology Clinical Database between the years 1980 and 2012; (3) could be linked to the Manitoba Health Insurance Registry; (4) had a valid date of diagnosis in the clinical database; (5) had a diagnosis recorded before the 16th birthday (i.e., the upper age limit for a diagnosis of JIA and JRA); (6) had continuous health insurance coverage in Manitoba from birth until the 16th birthday; and (7) were not missing their location of residence (i.e., postal code) at birth.

Children diagnosed with JIA, JRA, specific subtypes of these entities, or seronegative enthesopathy and arthropathy (SEA) syndrome by a pediatric rheumatologist as recorded in the Pediatric Rheumatology Clinical Database were classified as JA cases. Children recorded in the clinical database with other diagnoses were classified as non-cases. Since the non-cases were defined from only children referred to a pediatric rheumatologist, they do not represent the general population.

3.3 Study Variables

Tables 3.1 and 3.2 provide definitions for the four longitudinal disease markers included in the study: any JA-related visit, number of general practitioner visits, number of specialist visits, and presence of at least one hospitalization. The longitudinal disease markers were defined for 15 time periods beginning at baseline (i.e., birth). The first time period extended from birth to the second birthday (i.e., 0 and 1 years of age) and the remaining time periods were defined annually (i.e., 2, 3, 4, …, 15 years of age).

Table 3.1: Longitudinal Disease Marker Definitions

Longitudinal Disease Marker	**Variable Type**	**Data Source**	**Definition**
Any JA-related visit	Binary	Medical Claims DAD	At least one JA-related diagnosis code (Table 3.2) recorded in the Medical Claims database or Hospital Discharge Abstract Database within the corresponding time period
Number of general practitioner visits	Count	Medical Claims	Total number of ambulatory records in the Medical Claims database by a general practitioner with any diagnosis code within the corresponding time period
Number of specialist visits	Count	Medical Claims	Total number of ambulatory records in the Medical Claims database by a specialist physician with any diagnosis code within the corresponding time period
Hospitalization	Binary	DAD	At least one hospitalization recorded in the Hospital Discharge Abstract Database (excluding newborn records) with any diagnosis codes and the admission date within the corresponding time period

Table 3.2: Juvenile arthritis and related International Classification of Disease (ICD) codes

Description	ICD-9-CM[a]	ICD-10-CA[b]
Psoriatic arthritis	696	
Athropathy associated with other disorders classified elsewhere	713	
Rheumatoid arthritis	714	
Other and unspecified arthropathies	716	
Anklyosing spondylitis	720	M45
Seropositive rheumatoid arthritis		M05
Other rheumatoid arthritis		M06
Psoriatic and enteropathic arthropathies		M07
Juvenile arthritis		M08
Juvenile arthritis in diseases classified elsewhere		M09

[a] International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)
[b] International Classification of Diseases, Tenth Revision, Canada (ICD-10-CA)

Three covariates were included in the models to describe the longitudinal disease markers: sex, region of residence, and age. Sex was measured at baseline (i.e., birth) and was defined from the health insurance registry; it had two categories: male and female. Region of residence was measured at baseline (i.e., birth) and defined using postal codes recorded in the insurance registry; it had two categories: Winnipeg and Non-Winnipeg. Age was a time-varying continuous variable that was defined at the beginning of each time period (i.e., 0, 2, 3,…, 15 years).

Additional variables were used to describe the study cohort: age at diagnosis, period of diagnosis, and income quintile. Age at diagnosis was measured at the date of diagnosis recorded in the clinical database; it was categorized as: 1-5, 6-10, and 11-15 years. Period of diagnosis was also measured at the date of diagnosis recorded in the clinical database and was categorized into three 10-year periods: 1983-1992, 1993-2002, and 2003-2012. Income quintile, an area-level measure of socio-economic status based on total household income from Statistics Canada Census data, was measured at baseline (i.e., birth) by assigning individuals to income quintiles based on postal codes recorded in the insurance registry. Q1 represented the lowest income quintile and Q5 represented the highest income quintile. The quintiles were calculated separately for urban and rural areas, but were then combined. Due to small numbers, the Not Found (NF; i.e., missing) category was combined with Q1.

3.4 Dynamic Classification

The dynamic longitudinal discriminant analysis (LoDA) methods developed by Hughes et al. (2017) and Hughes, Komárek, et al. (2018), which use a multivariate generalized linear mixed model (MGLMM), were adapted for retrospective longitudinal administrative health data. Suppose that measurements for $R \geq 1$ longitudinal disease markers are made over time for N individuals and that each individual belongs to one of G groups (e.g., disease case or non-case). This group membership is represented by the random variable $U \in \{0, \dots, G-1\}$. Let $\mathbf{Y}_{i,r} = \left(Y_{i,r,1}, \dots, Y_{i,r,n_r}\right)$ denote the repeated measurements of the rth longitudinal disease marker for the ith individual observed at time points $\mathbf{t}_{i,r} = \left(t_{i,r,1}, \dots, t_{i,r,n_r}\right), t_{i,r,1} < \cdots < t_{i,r,n_r}$. It is not required that each longitudinal disease marker be measured at the same number or spacing of time points, nor is it required that each individual have the same number of measurements or measurement schedule. Also, let $\mathbf{v}_{i,r,1}, \dots, \mathbf{v}_{i,r,n_r} \in \mathbb{R}^{p_r}$ be vectors of baseline (e.g., sex, race/ethnicity) and possible time-varying (e.g., age, weight) covariates that may explain the evolution of the longitudinal disease markers. Therefore, complete information on the longitudinal disease markers, measurement times, and covariates for the ith individual is denoted by $\mathbb{Y}_i = \left(\mathbf{Y}_{i,1}, \dots, \mathbf{Y}_{i,R}\right)$ and $\mathcal{C}_i = \left(\mathbf{t}_{i,1}, \dots, \mathbf{t}_{i,R}, \mathbf{v}_{i,1,1}, \dots, \mathbf{v}_{i,R,n_R}\ \right)$.

In this study, $G = 2$, where the groups were defined as JA cases ($U = 1$) and non-cases ($U = 0$). The status of the individuals (i.e., JA case or non-case) was defined using the clinical database. The time points $\mathbf{t}_{i,r}$ were defined using 15 time periods starting at baseline (i.e., birth), where the first time period was from birth to the second birthday and the following time periods were one-year time periods (i.e., 2, 3, 4, …, 15 years of age). Thus, the time and spacing between measurements were the same for each individual, as well as for each of the longitudinal disease markers.

3.4.1 Multivariate Generalized Linear Mixed Model (MGLMM)

Separate MGLMMs were fit to each group using all longitudinal data. To fit the models, the group status of each individual must be known in the validation data source. Given the ith individual's group status $U_i = g$ and latent random effects vector $\mathbf{b}_i = (\mathbf{b}_{i,1}, \dots, \mathbf{b}_{i,R})$, the jth repeated measurement $Y_{i,r,j}$ $(j = 1, \dots, n_r)$ of the rth longitudinal disease marker $(r = 1, \dots, R)$ is assumed to follow a distribution from an exponential family (e.g., normal, binomial, Poisson) with a dispersion parameter ϕ_r^g. The expectation is given by

$$h_r^{-1}\{\mathrm{E}(Y_{i,r,j} | \mathbf{b}_i, U_i = g)\} = \mathbf{x}_{i,r,j}^{g^{\mathrm{T}}} \boldsymbol{\alpha}_r^g + \mathbf{z}_{i,r,j}^{g^{\mathrm{T}}} \mathbf{b}_{i,r}, \tag{3.1}$$

where h_r^{-1} is a chosen link function for the rth longitudinal disease marker, $\mathbf{x}_{i,r,j}^g = \mathbf{x}_{i,r,j}^g(\mathcal{C}_i)$ and $\mathbf{z}_{i,r,j}^g = \mathbf{z}_{i,r,j}^g(\mathcal{C}_i)$ are vectors of covariates derived from the information in $\mathcal{C}_i$, and $\boldsymbol{\alpha}_r^g$ are the unknown fixed effects parameters. The dispersion parameter ϕ_r^g is either known or unknown depending on the distribution of the rth disease marker (Hughes, Komárek, et al., 2018).

The MGLMM defined in (3.1) assumes the following: (1) the measurements for the rth longitudinal disease marker for the ith individual, $Y_{i,1,1}, \dots, Y_{i,R,n_R}$, are conditionally independent given their random effects vector $\mathbf{b}_i$; (2) the random vectors of the longitudinal disease markers, $\mathbb{Y}_i = (\mathbf{Y}_{i,1}, \dots, \mathbf{Y}_{i,R})$, are independent between individuals; and (3) the random effects vectors $\mathbf{b}_i$ are independent between individuals.

The random effects $\mathbf{b}_i = (\mathbf{b}_{i,1}, \dots, \mathbf{b}_{i,R})$, which can include both random intercepts and slopes, are included in (3.1) to account for the possible correlation between the repeated observations of the same longitudinal disease marker and are jointly distributed to model the possible correlation between the different longitudinal disease markers measured on the ith individual. The following normal mixture distribution is assumed for the random effects to robustify the model against possible misspecification of the random effects distribution

$$\mathbf{b}_i \mid U_i = g \sim \sum_{k=1}^{K^g} w_k^g \mathcal{MVN}\left(\boldsymbol{\mu}_k^g, \mathbb{D}_k^g\right), \tag{3.2}$$

where $\mathcal{MVN}(\boldsymbol{\mu}, \mathbb{D})$ denotes the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\mathbb{D}$. K^g is the number of mixture components for the gth group and is assumed to be known (Hughes, Komárek, et al., 2018). Unknown parameters of (3.2) for the gth group include the mixture weights $\mathbf{w}^g = \left(w_1^g, \dots, w_{K^g}^g\right)$ $(0 < w_k^g < 1, k = 1, \dots, K^g, \sum_{k=1}^{K^g} w_k^g = 1)$, mixture means $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{K^g}$, and mixture covariance matrices $\mathbb{D}_1, \dots, \mathbb{D}_{K^g}$.

To fit the MGLMM, the fixed effects coefficients $\boldsymbol{\psi}^g \coloneqq \left(\alpha_1^g, \dots, \alpha_R^g, \phi_1^g, \dots, \phi_R^g\right)$ and the mixture parameters $\boldsymbol{\theta}^g \coloneqq (\mathbf{w}^g, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{K^g}, \mathbb{D}_1, \dots, \mathbb{D}_{K^g})$ are estimated. Let $\mathbf{y}_{i,1}, \dots, \mathbf{y}_{i,R}$ be observed values of the longitudinal disease markers $\mathbb{Y}_i$ for the ith individual from group g, then the marginal density $f_g^{marg}(\cdot\,; \boldsymbol{\psi}^g, \boldsymbol{\theta}^g, \mathcal{C}_i)$ is defined as

$$f_g^{marg}\left(\mathbf{y}_{i,1}, \dots, \mathbf{y}_{i,R}; \boldsymbol{\psi}^g, \boldsymbol{\theta}^g, \mathcal{C}_i\right) = \int f_g^{cond}\left(\mathbf{y}_{i,1}, \dots, \mathbf{y}_{i,R} \middle| \mathbf{b}_i; \boldsymbol{\psi}^g, \mathcal{C}_i\right) f_g^{ranef}(\mathbf{b}_i; \boldsymbol{\theta}^g) \mathrm{d}\mathbf{b}_i \tag{3.3}$$

where $f_g^{cond}(\cdot\,|\mathbf{b}_i; \boldsymbol{\psi}^g, \mathcal{C}_i)$ is the conditional density of the observed longitudinal disease markers given the random effects and $f_g^{ranef}(\mathbf{b}_i; \boldsymbol{\theta}^g)$ is the density of the random effects. The conditional density $f_g^{cond}(\cdot\,|\mathbf{b}_i; \boldsymbol{\psi}^g, \mathcal{C}_i)$ in (3.3) is defined as

$$f_g^{cond}\left(\mathbf{y}_{i,1}, \dots, \mathbf{y}_{i,R} \middle| \mathbf{b}_i; \boldsymbol{\psi}^g, \mathcal{C}_i\right) = \prod_{r=1}^{R} \prod_{j=1}^{n_{i,r}} p_r\left(y_{i,r,j} | \mathbf{b}_i; \boldsymbol{\psi}^g, \mathcal{C}_i\right) \tag{3.4}$$

where $p_r(\cdot\,|\mathbf{b}_i;\boldsymbol{\psi}^g,\mathcal{C}_i)$ is the density assumed for the rth longitudinal disease marker. The density of the random effects $f_g^{ranef}(\mathbf{b}_i;\boldsymbol{\theta}^g)$ in (3.3) is defined as

$$f_g^{ranef}(\mathbf{b}_i;\boldsymbol{\theta}^g)=\sum_{k=1}^{K^g} w_k^g\varphi\left(\mathbf{b}_i;\boldsymbol{\mu}_k^g,\mathbb{D}_k^g\right) \tag{3.5}$$

where $\varphi(\cdot\,;\boldsymbol{\mu},\mathbb{D})$ denotes a density of the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbb{D}$. The covariance matrix $\mathbb{D}$ is unstructured.

Hughes et al. (2017) and Hughes, Komárek, et al. (2018) used a Markov Chain Monte Carlo (MCMC)-based Bayesian estimation approach developed by Komárek and Komárková (2013) to estimate the unknown parameters of the MGLMM. The MCMC approach uses a block Gibbs algorithm to generate a sample of size M, $\mathcal{S}_g=\left\{\left(\boldsymbol{\psi}^{g,(m)},\boldsymbol{\theta}^{g,(m)}\right): m=1,\dots,M\right\}$, from the following posterior distribution

$$p\left(\boldsymbol{\psi}^g,\boldsymbol{\theta}^g\middle|\mathbf{y}_g\right)\propto L_g(\boldsymbol{\psi}^g,\boldsymbol{\theta}^g)p(\boldsymbol{\psi}^g,\boldsymbol{\theta}^g) \tag{3.6}$$

where $L_g(\boldsymbol{\psi}^g,\boldsymbol{\theta}^g)$ is the likelihood of the MGLMM for group g and $p(\boldsymbol{\psi}^g,\boldsymbol{\theta}^g)$ is the prior distribution of the model parameters, which is specified to be weakly informative. Assuming independence between individuals, the likelihood $L_g(\boldsymbol{\psi}^g,\boldsymbol{\theta}^g)$ is defined as

$$L_g(\boldsymbol{\psi}^g,\boldsymbol{\theta}^g)=\prod_{i:\,u_i=g} f_g^{marg}\left(\mathbf{y}_{i,1},\dots,\mathbf{y}_{i,R};\boldsymbol{\psi}^g,\boldsymbol{\theta}^g,\mathcal{C}_i\right). \tag{3.7}$$

More detailed information about the MCMC sampling procedure and the specification of the prior distribution can be found in the appendices of Komárek and Komárková (2013). The described MCMC methods can be implemented using the R package mixAK (Komárek & Komárková, 2014).

The number of mixture components (K^g) for the random effects distribution must also be selected to fit the MGLMMs. It is not necessary for K^g to be equal between the groups. The optimal number of mixture components can be determined by assessing model fit using the penalized expected deviance (PED), a fit statistic for mixture models (Plummer, 2008). Lower values for the PED indicate better model fit. For simplicity, K is taken to be 1 for both JA cases and non-cases in this study, meaning the random effects followed a multivariate normal distribution.

3.4.2 Group Membership Probabilities

After estimating the parameters of the MGLMMs, the aim of the discriminant analysis is to classify an individual using their longitudinal history. Given the individual's longitudinal and explanatory variable data, as well as the estimated model parameters for the MGLMMs, Bayes theorem gives the probability that the individual belongs to group g as

$$\mathcal{P}_g = \frac{\pi_g f_g}{\sum_{\tilde{g}=0}^{G-1} \pi_{\tilde{g}} f_{\tilde{g}}} \quad g = 0, \dots, G-1 \tag{3.8}$$

where f_g is a predictive density of the individual's observed longitudinal disease markers given the group and model parameters and $\pi_g = P(U = g)$ are prior probabilities of belonging to each group ($0 < \pi_g < 1, \sum_{g=0}^{G-1} \pi_g = 1$) (Hughes, El Saeiti, et al., 2018). Typically, the prior group probabilities are defined as the prevalence of the groups in the study population.

In this study, the cohort is a referral population and may not represent the general population; therefore, the prevalence of JA in the Manitoba population could be used to define the prior group probabilities, but would not be accurate. Initially, naïve prior group probabilities of 0.50 for both JA cases and non-cases were used. This was similar to the distribution in the study cohort, where 48% of the cohort were JA cases and 52% were non-cases. Additional hypothetical prior probabilities of being a JA case of 0.30, 0.40, 0.60, and 0.70 were used to determine whether changing the prior group probabilities had an impact on classification accuracy.

The predictive density f_g in equation 3.8 can be specified using one of three prediction approaches: marginal, conditional, and random effects (Hughes, El Saeiti, et al., 2018). For the marginal prediction approach, the predictive density f_g is taken as the marginal density $f_g^{marg}(y_{i,1}, \dots, y_{i,R}; \boldsymbol{\psi}^g, \boldsymbol{\theta}^g, C_i)$ defined in (3.3). The aim of the marginal approach is to compare an individual's observed longitudinal history to the mean longitudinal profiles of each of the groups. The individual is classified to the group that has the most similar longitudinal profile. For the conditional prediction approach, the predictive density f_g is taken as the conditional density $f_g^{cond}(y_{i,1}, \dots, y_{i,R} | \mathbf{b}_i^g; \boldsymbol{\psi}^g)$ defined in (3.4). The conditional approach compares the conditional longitudinal profile of an individual with the mean longitudinal profiles of a subset of individuals with similar random effects in each group. For the random effects prediction approach, the predictive density f_g is taken as the random effects density $f_g^{ranef}(\mathbf{b}_i^g; \boldsymbol{\theta}^g)$ defined in (3.5). The focus of the random effects approach is the patient-specific evolution of the longitudinal disease markers. The group membership probabilities for the marginal, conditional, and random effects prediction approaches are approximated with their posterior means calculated using the M samples generated from the MCMC scheme (Komárek & Komárková, 2013).

In general, the accuracy of the prediction approaches depends on the data being used and must be evaluated in the process of building and testing the model. Hughes, El Saeiti, et al. (2018) conducted a simulation study to explore situations in which the three approaches work best. The marginal approach is expected to provide the best prediction accuracy when there are noticeable differences in the mean longitudinal profiles between the groups. If instead, the main difference between the groups is the variability around the mean profiles, the random effects approach is expected to provide the best prediction accuracy. However, if there is large measurement error or there are not enough repeated measurements per individual to precisely estimate the random effects, the marginal approach may provide more accurate predictions. The authors concluded that the conditional approach added little value, as they were unable to identify a scenario where it outperformed both the marginal and random effects approaches simultaneously.

3.4.3 Dynamic Classification Procedure

For an individual, let $\mathbf{Y}_r(t)$ be a subvector of $\mathbf{Y}_r$ that includes the repeated measurements for the rth longitudinal disease marker accrued by time t, $\mathbb{Y}(t) = \big(\mathbf{Y}_1(t), \dots, \mathbf{Y}_R(t)\big)$ be the observed values of all longitudinal disease markers accrued by time t, and $\mathcal{C}(t)$ denote the covariate values accrued by time t. Dynamic classification uses an individual's longitudinal history $\mathbb{Y}(t)$ and covariate information $\mathcal{C}(t)$ accrued by a given time point t to predict their value of the group variable U, as soon as possible. The individual's group membership probabilities are calculated at each consecutive time point using their updated longitudinal information until they can be confidently classified into one of the G groups based on a chosen allocation scheme. Once an individual is classified into one of the G groups, their group membership probabilities are not updated further (i.e., their classification does not change) (Hughes et al., 2017; Hughes, Komárek, et al., 2018).

In the prospective setting, a time point t is defined as the visit time of an individual, meaning their longitudinal information and group membership probabilities are updated each time they visit their health care provider. In this study's retrospective setting, time was defined using 15 time periods from baseline (i.e., birth), meaning an individual's longitudinal information and corresponding group membership probabilities were updated after every time period.

Hughes et al. (2017) developed an allocation scheme that incorporates the credible intervals (CrIs) of the group membership probabilities to account for the variability between individuals in the uncertainty of the group membership probabilities. The CrI allocation scheme increased the positive predictive value (PPV) of the discriminant analysis without sacrificing the sensitivity, specificity, and probability of correct classification (PCC). The $(1-\alpha)100\%$ CrI for the group membership probability $\mathcal{P}_g(t)$ defined in (3.9) calculated at time t is denoted by $\left(\mathcal{P}_g^{LOW}(t), \mathcal{P}_g^{UPP}(t)\right)$ if

$$\mathrm{P}\left[\mathcal{P}_g(t) \in \left(\mathcal{P}_g^{LOW}(t), \mathcal{P}_g^{UPP}(t)\right)\right] = 1-\alpha. \quad (3.9)$$

Equation (3.9) does not uniquely determine the CrI, but states that the interval captures $1-\alpha$ of the probability mass of the posterior distribution of $\mathcal{P}_g(t)$. If the posterior distribution is unimodal, then the shortest CrI is called the highest posterior density (HPD) CrI. If the posterior distribution of $\mathcal{P}_g(t)$ is not unimodal, equal-tail CrIs can be used (Hughes et al., 2017). Let $\left(\mathcal{P}_g^{LOW}(t), \mathcal{P}_g^{UPP}(t)\right)$ denoted the $(1-\alpha)100\%$ HPD CrI for $\mathcal{P}_g(t)$. The CrIs are calculated using the MCMC samples.

In the two-group case (i.e., $G = 2$), as in the JA application, the probability of belonging to group 1 (i.e., JA case) is used in the allocation scheme. Let $\mathcal{P}_1(t)$ denote the probability of belonging to group 1 (i.e., the probability of being a JA case) and $\left(\mathcal{P}_1^{LOW}(t), \mathcal{P}_1^{UPP}(t)\right)$ be the corresponding $(1-\alpha)100\%$ CrI. Adapted from Hughes et al. (2017), the steps of the dynamic CrI classification scheme to predict an individual's group membership at time t, $\widehat{U}(t)$, based on a cut-off value $c \in (0,1)$ are:

1. Calculate the $(1-\alpha)100\%$ CrI $\left(\mathcal{P}_1^{LOW}(t), \mathcal{P}_1^{UPP}(t)\right)$ for the group membership probability $\mathcal{P}_1(t)$.
2. If $\mathcal{P}_1^{LOW}(t) > c$, assign individual to group 1 (i.e., JA case), $\widehat{U}(t) = 1$.
3. If $\mathcal{P}_1^{UPP}(t) < c$, assign individual to group 0 (i.e., JA non-case), $\widehat{U}(t) = 0$.
4. If $\mathcal{P}_1^{LOW}(t) \leq c \leq \mathcal{P}_1^{UPP}(t)$, the individual is left unclassified, $\widehat{U}(t) =$ NA.
5. If the individual remains unclassified at time t, update the probability of belonging to group 1 and its corresponding CrI calculated in Step 1 at the next time point, and then follow Steps 2-5.
6. Continue until all individuals are classified or until observations from all time points are used and no further classifications can be achieved.

An individual is only classified into groups 1 or 0 (i.e., JA case or non-case) if the cut-off c falls outside of the $(1-\alpha)100\%$ CrI for $\mathcal{P}_1(t)$ (Steps 2 and 3). If c falls within the $(1-\alpha)100\%$ CrI for $\mathcal{P}_1(t)$, then the individual is not classified and is referred to the next time point where further information is available. In this study, 99% CrIs were used to make classification decisions, as they were previously shown to perform the best (Hughes, Bonnett, et al., 2018; Hughes et al., 2017).

The value of the cut-off c must be chosen *a priori* by the investigator. The point nearest to the top left corner of the ROC curve, which minimizes $d^2 = (1 - Sensitivity)^2 + (1 - Specificity)^2$, was chosen as the cut-off for this study. This cut-off has been previously shown in the dynamic classification setting to be optimal for balancing sensitivity and specificity (Hughes et al., 2017; Morrell et al., 2012).

3.5 Statistical Analysis

Descriptive statistics of the study cohort were produced separately for JA cases and non-cases using frequencies and percentages. Linear trends in the longitudinal disease markers were described for JA cases and non-cases.

3.5.1 Model Development

MGLMMs were fit separately for JA cases and non-cases using all of the longitudinal information available for each individual within the groups. Bernoulli models with logit link functions were fit for the binary longitudinal disease markers (i.e., any JA-related visit and hospitalization) and Poisson models with log link functions were fit for the count longitudinal disease markers (i.e., number of general practitioner visits and number of specialist visits). Fixed effects covariates (i.e., sex, region of residence, and age) were included for each of the longitudinal disease markers, as well as a random intercept to allow the means of the longitudinal disease markers to vary across individuals. The following four combinations of longitudinal disease markers were fit to the longitudinal data:

1. **Model 1**: Any JA-related visit
2. **Model 2**: Any JA-related visit + Number of specialist visits
3. **Model 3**: Any JA-related visit + Number of specialist visits + Hospitalization
4. **Model 4**: Any JA-related visit + Number of specialist visits + Hospitalization + Number of general practitioner visits

These four models were compared based on their classification accuracy when used with the dynamic classification procedure. The models were fit using the MCMC algorithm developed by Komárek and Komárková (2013) and the parameters of the models were described using their posterior means and 95% CrIs.

The convergence of the MCMC algorithm was evaluated using the coda package in R. Two chains beginning at different positions were sampled from the MCMC algorithm. Trace plots, plots of the iteration number against the value of the MCMC sample, were produced to visually assess if the two chains covered to the same posterior distribution for the parameter(s) of interest. In addition, the Gelman-Rubin diagnostic was used to assess convergence (Gelman & Rubin, 1992). The Gelman-Rubin diagnostic uses the potential scale reduction factor (PSRF) to measure whether there is a significant difference between the variance within the two chains and the variance between the two chains. Large differences between the variances indicate non-convergence, meaning if the PSRF (and its upper confidence limit) are close to 1, it can be concluded that the chains have converged to the target posterior distribution (Gelman & Rubin, 1992). Finally, auto-correlation plots were produced to determine whether dependence existed between the MCMC samples.

3.5.2 Evaluation of Dynamic Classification

Five-fold cross-validation was used to evaluate the dynamic classification procedure. The study cohort was randomly split into five folds where each fold contained 20% of the cohort. For fold 1, the MGLMMs were fit using the data from folds 2-5. Then, the dynamic LoDA model was used to classify each individual in fold 1 as either a JA case or non-case, or left them unclassified. This was repeated for the remaining folds. A confusion matrix that includes an "Unclassified" classification category (Table 3.3) was created for each fold after the last time period using the optimal cut-off point (i.e., point nearest to the top left corner of the ROC curve), the classifications from the dynamic LoDA and the disease status from the clinical database. The confusion matrices for each fold were used to calculate classification accuracy measures (Table

3.4), which were then averaged over the five folds to evaluate the accuracy of the dynamic LoDA.

Classification accuracy was compared amongst the prediction approaches (i.e., marginal, conditional, and random effects), as well as between the four models. The best dynamic LoDA model was chosen to be the model with the highest AUC. Using the best model, classification accuracy was also compared between varying prior group probabilities. In addition, proportion unclassified, sensitivity, specificity, and PPV were calculated for the best model after each time period to demonstrate the evolution of classification accuracy over time.

Table 3.3: Confusion matrix for JA cases and non-cases

		Classification			
		JA Case	**JA Non-Case**	**Unclassified**	**Total**
Disease Status from Clinical Database	**JA Case**	A	B	C	A+B+C
	JA Non-Case	D	E	F	D+E+F
	Total	A+D	B+E	C+F	A+B+C+D+E+F

Table 3.4: Definitions of classification accuracy measures

Measure	Definition
Sensitivity (classified data)*	$Sensitivity^* = \frac{A}{A+B}$
Sensitivity	$Sensitivity = \frac{A}{A+B+C}$
Specificity (classified data)*	$Specificity^* = \frac{E}{D+E}$
Specificity	$Specificity = \frac{E}{D+E+F}$
Probability of Correct Classification (PCC; classified data)*	$PCC^* = \frac{A+E}{A+B+D+E}$
Probability of Correct Classification (PCC)	$PCC = \frac{A+E}{A+B+C+D+E+F}$
Area Under Curve (AUC)	Area under the receiver operating characteristic (ROC) curve
Positive Predictive Value (PPV)	$PPV = \frac{A}{A+D}$
Negative Predictive Value (NPV)	$NPV = \frac{E}{B+E}$
Proportion Unclassified	$Unclassified = \frac{C+F}{A+B+C+D+E+F}$
Mean Classification Time (years)	Classification time is defined as the number of years of data needed to make a classification as a JA case or non-case

*Measure is calculated without including the unclassified individuals in the denominator

3.5.3 Static Model-Based and Deterministic Case Definitions

A static model-based case definition and a validated deterministic case definition were also applied to the study cohort. Static LoDA classified individuals as JA cases or non-cases using their entire longitudinal history at the end of a specified period of time. The dynamic LoDA model with the highest AUC was used. Four static LoDA models were applied to the study cohort using the following periods of data: (1) birth to 6th birthday (i.e., 0-5 years of age), (2) birth to 11th birthday (i.e., 0-10 years of age), (3) birth to 16th birthday (i.e., 0-15 years of age); and (4) 14th to 16th birthdays (i.e., 14-15 years of age).

The JA case definition validated by Shiff et al. (2017), which was “1 or more hospitalizations or 2 or more diagnoses in 2 years by any provider 8 or more weeks apart using diagnosis codes for rheumatoid arthritis and ankylosing spondylitits (ICD-9-CM codes: 714, 720; ICD-10-CA codes: M05, M06, M08, M45)”, was also applied to the study cohort. The deterministic case definition was applied to data from two periods: (1) birth to 16th birthday (i.e., 0-15 years of age) and (2) 14th to 16th birthdays (i.e., 14-15 years of age).

Sensitivity, specificity, PCC, PPV, and NPV were produced for the static LoDA and deterministic case definitions and compared to the classification accuracy of the dynamic LoDA model with the highest AUC. In addition, sex-specific dynamic and static LoDA models were produced to determine whether increased classification accuracy could be achieved by fitting the LoDA models separately for males and females.

3.6 Ethics Approvals

Ethics approval for this study was provided by the University of Manitoba Health Research Ethics Board (HREB: HS22501). Data access was granted by the Health Information Privacy Committee (HIPC: 2018/2019-62) and the Winnipeg Regional Health Authority.

CHAPTER 4 – RESULTS

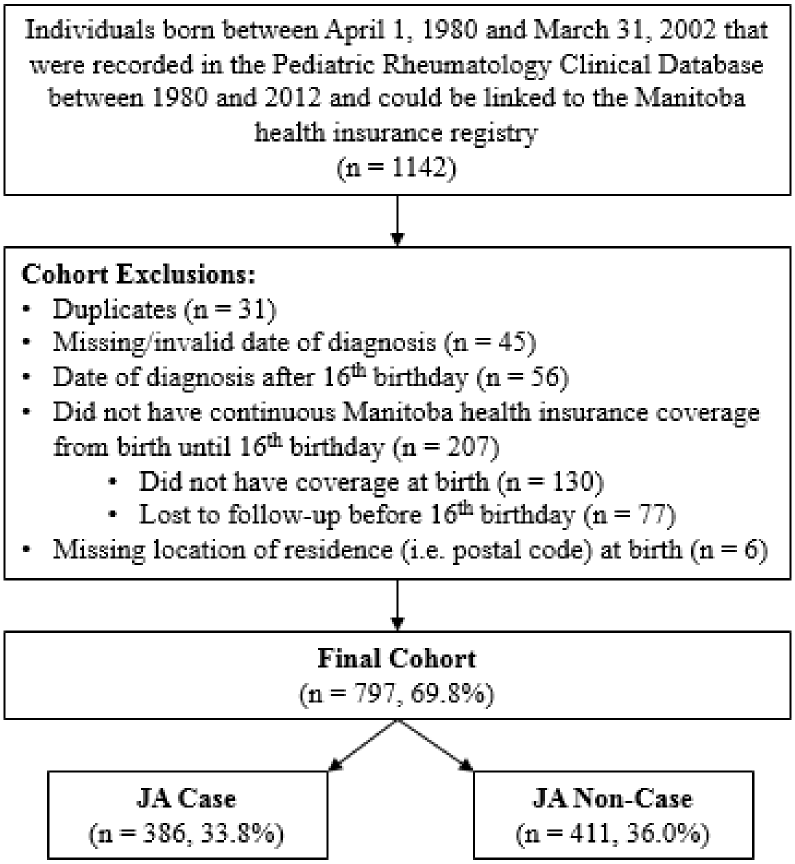
This chapter provides the results of the JA application in five sections. First, the study cohort is described. Second, the results of the MGLMMs are presented. Third, the classification accuracy of the dynamic LoDA models are summarized. Fourth, the classification accuracy of the static LoDA model and deterministic case definition are summarized and compared to the best dynamic LoDA model. Finally, the sex-specific results are provided.

4.1 Description of Study Cohort

A total of 1142 children were born between April 1, 1980 and March 31, 2002 and were recorded in the Pediatric Rheumatology Clinical Database in the years 1980 to 2012 (Figure 4.1). After applying the inclusion criteria, the study cohort included 797 children, of which 386 (48.4%) were JA cases and 411 (51.6%) were non-cases.

A total of 207 individuals were excluded from the final cohort because they did not have continuous Manitoba health insurance coverage from birth to 16th birthday. Almost two-thirds (62.8%) of these individuals did not have health insurance coverage at birth. Overall, close to one-fifth (16.9%) of the individuals who did not have coverage at birth gained coverage before the first birthday. Individuals who gained coverage before their first birthday were more likely to be a non-Winnipeg resident when registered. It is possible that some individuals may not have been registered at birth due to administrative error. The remaining individuals had health insurance coverage at birth but were lost to follow-up before their 16th birthday. Overall, three-quarters (76.6%) of these individuals left the province, 15.6% could not be located or were registered in error, and 7.8% were deceased.

Figure 4.1: Cohort Flowchart


Individuals born between April 1, 1980 and March 31, 2002 that were recorded in the Pediatric Rheumatology Clinical Database between 1980 and 2012 and could be linked to the Manitoba health insurance registry
(n = 1142)
Cohort Exclusions:
• Duplicates (n = 31)
• Missing/invalid date of diagnosis (n = 45)
• Date of diagnosis after 16th birthday (n = 56)
• Did not have continuous Manitoba health insurance coverage from birth until 16th birthday (n = 207)
• Did not have coverage at birth (n = 130)
• Lost to follow-up before 16th birthday (n = 77)
• Missing location of residence (i.e. postal code) at birth (n = 6)
Final Cohort
(n = 797, 69.8%)
JA Case
(n = 386, 33.8%)
JA Non-Case
(n = 411, 36.0%)

Characteristics of the study cohort stratified by JA cases and non-cases are provided in Table 4.1. Cohort members in both groups were more likely to be female, which was expected as pediatric rheumatic diseases are more common among girls (Petty et al., 2016; Shiff et al., 2014). The distribution of age at diagnosis recorded in the clinical database was slightly different between the groups. Non-cases were more likely to be diagnosed older, whereas cases were more likely to be diagnosed either in early childhood (i.e., 0-5 years of age) or early adolescence (i.e., 11-15 years of age). Period of diagnosis was also dissimilar between the two groups. Although the majority of cohort members in both groups were diagnosed between 1993 and 2002, almost 80% of the non-cases were diagnosed during those years compared to only half of the cases.

Additionally, only 1.5% of the non-cases were diagnosed between 1983 and 1992, compared to 21.5% of the cases. Cohort members in both groups were more likely to live in Winnipeg at birth and were generally evenly distributed across the income quintiles. However, there was a slight increase in the percentage of children in the lowest income quintile amongst the non-cases compared to the cases.

Table 4.1: Cohort characteristics, n (%)

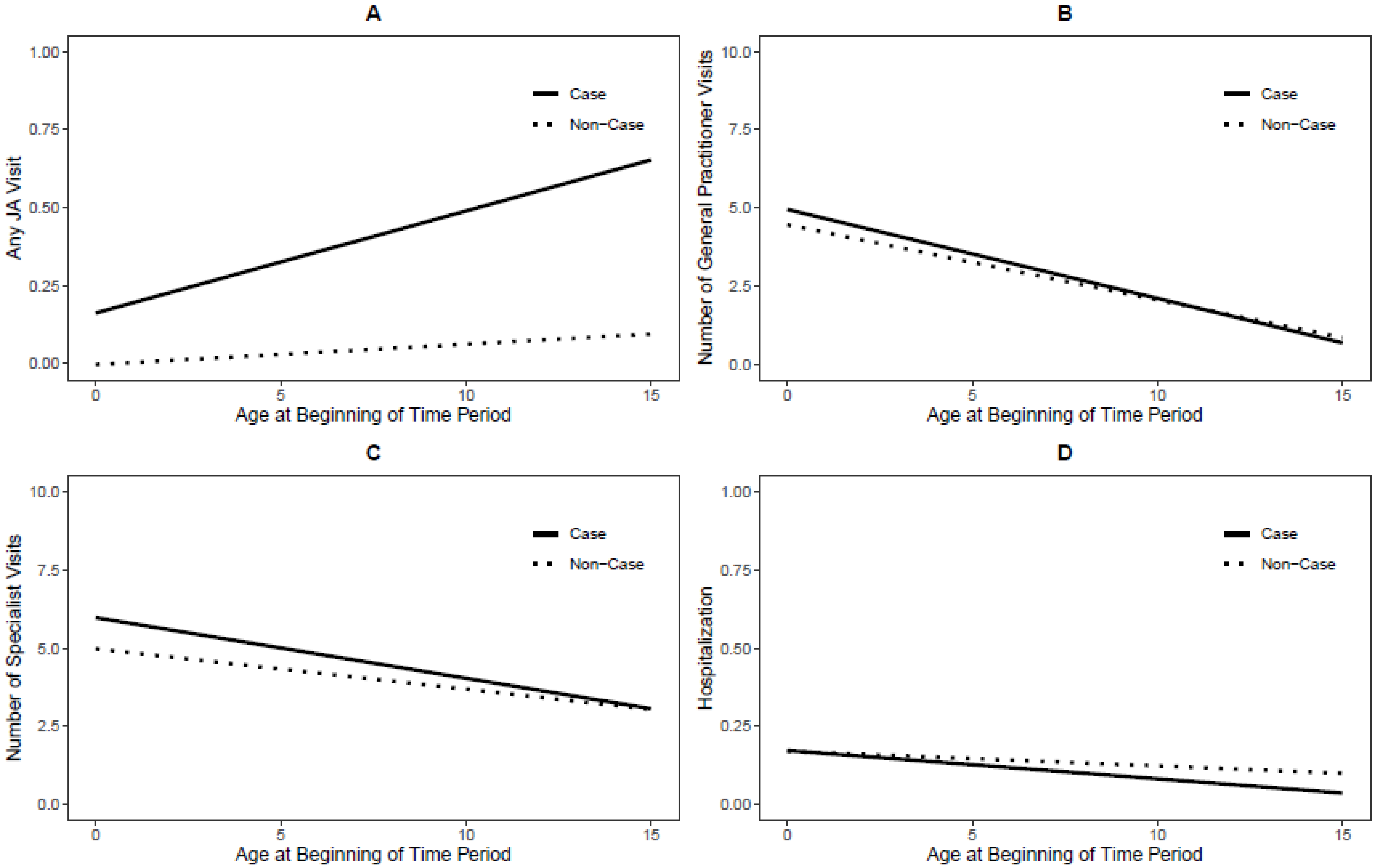
Variable	JA Cases (n = 386)	JA Non-Cases (n = 411)	P-value*
Sex			
Male	134 (34.7)	145 (35.3)	0.87
Female	252 (65.3)	266 (64.7)	
Age at Diagnosis (years)			
0–5	146 (37.8)	102 (24.8)	<0.01
6–10	82 (21.2)	118 (28.7)	
11–15	158 (40.9)	191 (46.5)	
Period of Diagnosis			
1983–1992	83 (21.5)	6 (1.5)	<0.01
1993–2002	192 (49.7)	326 (79.3)	
2003–2012	111 (28.8)	79 (19.2)	
Region of Residence at Birth			
Winnipeg	218 (56.5)	238 (57.9)	0.68
Non-Winnipeg	168 (43.5)	173 (42.1)	
Income Quintile at Birth			
Q1–Lowest/Not Found	83 (21.5)	102 (24.8)	0.48
Q2	68 (17.6)	77 (18.7)	
Q3	80 (20.7)	67 (16.3)	
Q4	81 (21.0)	91 (22.1)	
Q5–Highest	74 (19.2)	74 (18.0)	

*P-value was generated from chi-square test of independence

Figure 4.2 provides the linear trends over time for each of the longitudinal disease markers for cases and non-cases. As expected, JA cases were more likely to have a JA visit than non-cases at all ages. This likelihood increased consistently amongst cases over time. For non-cases, the likelihood of having a JA visit was low at all ages, but did increase over time. The trend for the number of visits with a general practitioner was very similar between cases and non-cases. For both groups, the number of visits with a general practitioner was highest early in life and then decreased steadily until age 15. Similarly, the number of visits with a specialist physician was highest for both groups early in life and then decreased until age 15. For the non-cases, the number of specialist visits was lower than for the cases until around age 10. For both groups, the likelihood of hospitalization was slightly higher earlier in life than later in life, but was low across all ages. The likelihood of hospitalization was slightly higher for non-cases than for cases starting around age 8.

Figure 4.2 shows that the longitudinal disease markers varied by age; and Figures A.1 to A.4 (Appendix A) illustrate how the linear trends for the longitudinal disease markers vary by sex and region of residence at birth. For cases, females were more likely to have a JA visit compared to males across all ages. For both JA cases and non-cases, the number of general practitioner visits were higher for non-Winnipeg residents compared to Winnipeg residents until around age 10. Conversely, the number of specialist visits were higher for Winnipeg residents compared to non-Winnipeg residents. For JA cases, the number of specialist visits were higher for females compared to males. For both JA cases and non-cases, the likelihood of hospitalization was slightly higher for non-Winnipeg residents compared to Winnipeg residents, as well as for males compared to females.

Figure 4.2: Linear trends stratified by JA cases and non-cases for longitudinal disease markers, A) Any JA visit, B) Number of general practitioner visits, C) Number of specialist visits, and D) Hospitalization



4.2 Multivariate Generalized Linear Mixed Models

Models 1to 4 were fit to all longitudinal data for the JA cases and non-cases. A summary of the estimated model parameters for each of the models is provided in Table 4.2. The fixed effect terms were determined to be statistically significant if 0 did not fall within the corresponding 95% HPD credible interval. The estimated model parameters for the fixed effects and random intercepts were similar for each of the four models, so only the results of model 4, which included all four longitudinal disease markers, are described in detail.

For JA cases, sex and age were significantly associated with having a JA visit. Females were more likely to have a JA visit and increased age was associated with an increased log-odds of having a JA visit. For non-cases, region of residence and age were statistically significant. Children living in Winnipeg at birth were less likely to have a JA-related visit compared those living outside of Winnipeg, and similar to the cases, increased age was associated with an increased log-odds of having a JA-related visit. The expected value of the random intercept for non-cases was -4.74, which was 62.6% lower than the expected valued of random intercept for cases of -1.77.

For JA cases, sex, region of residence, and age were significantly associated with the number of visits with a specialist physician. Females, those living in Winnipeg at birth, and younger age were associated with increased number of specialist visits. For non-cases, region of residence and age were statistically significant. Similar to the cases, non-cases that lived in Winnipeg at birth and younger ages were associated with increased number of specialist visits. The expected values of the random intercepts for cases and non-cases were 1.26 and 0.91, respectively.

Table 4.2: Posterior means and 95% highest posterior density credible intervals (CrI) for the fixed effects and random intercepts in models 1–4 for JA cases and non-cases

	Model 1[a]		**Model 2[b]**		**Model 3[c]**		**Model 4[d]**	
	Cases	**Non-Cases**	**Cases**	**Non-Cases**	**Cases**	**Non-Cases**	**Cases**	**Non-Cases**
Any JA-related visit								
Male (vs Female)	**-0.60** **(-0.97,-0.24)**	-0.07 (-0.44,0.28)	**-0.60** **(-0.96,-0.24)**	-0.08 (-0.44,0.28)	**-0.60** **(-0.95,-0.22)**	-0.07 (-0.44,0.28)	**-0.59** **(-0.95,-0.22)**	-0.07 (-0.43,0.27)
Winnipeg (vs Non-Winnipeg)	-0.24 (-0.59,0.11)	**-0.57** **(-0.91,-0.22)**	-0.24 (-0.57,0.12)	**-0.54** **(-0.87,-0.20)**	-0.24 (-0.60,0.11)	**-0.54** **(-0.88,-0.20)**	-0.24 (-0.60,0.11)	**-0.53** **(-0.86,-0.18)**
Age	**0.21** **(0.19,0.22)**	**0.17** **(0.14,0.20)**	**0.21** **(0.19,0.22)**	**0.17** **(0.14,0.20)**	**0.21** **(0.19,0.22)**	**0.17** **(0.14,0.20)**	**0.21** **(0.19,0.22)**	**0.17** **(0.14,0.20)**
E(Intercept)	-1.76 (-2.09,-1.44)	-4.75 (-5.22,-4.28)	-1.77 (-2.10,-1.44)	-4.74 (-5.22,-4.28)	-1.78 (-2.10,-1.44)	-4.74 (-5.22,-4.27)	-1.77 (-2.11,-1.46)	-4.74 (-5.21,-4.28)
SD(Intercept)	1.59 (1.44,1.75)	1.06 (0.86,1.27)	1.59 (1.44,1.74)	1.04 (0.84,1.24)	1.59 (1.44,1.75)	1.04 (0.83,1.24)	1.60 (1.44,1.75)	1.03 (0.84,1.23)
Number of specialist visits								
Male (vs Female)			**-0.19** **(-0.36,-0.02)**	-0.13 (-0.32,0.06)	**-0.19** **(-0.36,-0.01)**	-0.13 (-0.31,0.07)	**-0.19** **(-0.37,-0.02)**	-0.13 (-0.32,0.06)
Winnipeg (vs Non-Winnipeg)			**0.55** **(0.38,0.72)**	**0.64** **(0.44,0.82)**	**0.55** **(0.38,0.71)**	**0.63** **(0.45,0.82)**	**0.55** **(0.38,0.71)**	**0.63** **(0.45,0.81)**
Age			**-0.04** **(-0.05,-0.04)**	**-0.03** **(-0.04,-0.03)**	**-0.04** **(-0.05,-0.04)**	**-0.03** **(-0.04,-0.03)**	**-0.04** **(-0.05,-0.04)**	**-0.03** **(-0.04,-0.03)**
E(Intercept)			1.26 (1.11,1.40)	0.90 (0.74,1.06)	1.26 (1.11,1.40)	0.91 (0.74,1.06)	1.26 (1.12,1.40)	0.91 (0.74,1.07)
SD(Intercept)			0.81 (0.75,0.87)	0.93 (0.86,1.00)	0.80 (0.75,0.87)	0.92 (0.85,0.99)	0.80 (0.74,0.86)	0.92 (0.85,0.99)
Hospitalization								
Male (vs Female)					0.23 (-0.05,0.49)	0.01 (-0.22,0.25)	0.24 (-0.03,0.52)	0.02 (-0.23,0.25)

Winnipeg (vs Non-Winnipeg)	**-0.57** **(-0.82,-0.30)**	**-0.58** **(-0.83,-0.36)**	**-0.57** **(-0.83,-0.32)**	**-0.58** **(-0.80,-0.34)**
Age	**-0.11** **(-0.13,-0.09)**	**-0.04** **(-0.06,-0.03)**	**-0.11** **(-0.13,-0.09)**	**-0.04** **(-0.06,-0.03)**
E(Intercept)	-1.46 (-1.72,-1.20)	-1.46 (-1.70,-1.24)	-1.46 (-1.72,-1.20)	-1.47 (-1.71,-1.24)
SD(Intercept)	0.87 (0.73,1.02)	0.85 (0.73,0.96)	0.87 (0.73,1.01)	0.84 (0.73,0.96)
Number of general practitioner visits				
Male (vs Female)			0.03 (-0.16,0.24)	-0.02 (-0.22,0.19)
Winnipeg (vs Non-Winnipeg)			**-0.50** **(-0.68,-0.30)**	**-0.46** **(-0.66,-0.25)**
Age			**-0.11** **(-0.11,-0.10)**	**-0.10** **(-0.10,-0.09)**
E(Intercept)			1.62 (1.46,1.79)	1.45 (1.28,1.63)
SD(Intercept)			0.94 (0.87,1.02)	1.00 (0.92,1.07)

Note: Values in bold face font indicate fixed effect estimate was statistically significant at α = 0.05

[a] Model 1: Any JA-related visit

[b] Model 2: Any JA-related visit + Number of specialist visits

[c] Model 3: Any JA-related visit + Number of specialist visits + Hospitalization

[d] Model 4: Any JA-related visit + Number of specialist visits + Hospitalization + Number of general practitioner visits

For both JA cases and non-cases, region of residence and age were significantly associated with hospitalization. Children living outside of Winnipeg at birth were more likely to be hospitalized than those living in Winnipeg. Increased age was associated with decreased log-odds of hospitalization for both cases and non-cases, but with a higher magnitude for cases. The expected values of the random intercepts were similar for cases and non-cases at -1.46 and -1.47, respectively.

For the number of general practitioner visits, region of residence and age were statistically significant for both JA cases and non-cases. The estimated fixed effects parameters were similar for cases and non-cases; younger children and those living outside of Winnipeg at birth had a greater number of general practitioner visits than older children and those living in Winnipeg, respectively. The expected values of the random intercepts for cases and non-cases, were 1.62 and 1.45, respectively.

For each of the four models, visual assessment of the trace plots indicated that convergence was reach after the 1000th iteration. A total of 10 000 MCMC samples were drawn from the Gibbs sampler. The Gelman-Rubin diagnostic was used to ensure convergence to the target posterior distribution was reached. The upper confidence limits of the PSRF were less than 1.02 for all parameters, suggesting convergence was reached with 10 000 samples. Autocorrelation plots indicated that the MCMC samples were correlated, thus, 1:100 thinning was applied. In summary, after determining convergence was reached by the 1000th iteration, the first 1000 samples were discarded as “burn-in” and the remaining 9000 samples were used for inference. Figures B.1 to B.20 (Appendix B) provide the trace plots, density plots, Gelman-Rubin-Brooks diagnostic plots and autocorrelation plots for the posterior distributions of the fixed effects and random intercept parameters.

4.3 Dynamic Classification

Table 4.3 provides a summary of the classification accuracy for dynamic LoDA using each of the four models with naïve prior probabilities (i.e., prior probabilities of 0.50 for both cases and non-cases). The marginal and random effects predication approaches are shown. The conditional prediction approach performed similarly to the marginal approach and its results are provided in Appendix C.

For each model, the random effects prediction approach had higher classification accuracy on every measure compared to the marginal approach. However, the random effects prediction approach left 40% (model 2) to 67% (model 4) more individuals unclassified after the last time period compared to the marginal approach. In addition, for every model, the random effects prediction had a higher mean classification time than the marginal approach, meaning, on average, it required more years of data to make a classification. On average, 3.61 (model 2) to 6.44 (model 1) additional years of data were needed to classify an individual as a JA case or non-case using the random effects approach.

For the marginal approach, model 4 performs best on all classification accuracy measures, except for specificity, which is highest for model 1 at 0.74. No individuals were left unclassified at the end of the time periods using model 1. The proportion of unclassified individuals for models 2 to 4 were 0.05, 0.08, and 0.09, respectively. Mean classification time was considerably lower for model 1 at 2.77 years and incrementally increased for models 2 to 4 at 5.15, 6.15, and 7.07 years, respectively.

Table 4.3: Summary of the classification accuracy for dynamic LoDA using models 1–4 with the marginal and random effects prediction approaches and naïve prior probabilities

	Model 1[a]		**Model 2[b]**		**Model 3[c]**		**Model 4[d]**	
	Marginal	**Random Effects**	**Marginal**	**Random Effects**	**Marginal**	**Random Effects**	**Marginal**	**Random Effects**
Cut-off	0.44	0.35	0.43	0.42	0.44	0.39	0.48	0.43
Sensitivity (classified)*	0.46	0.71	0.62	0.72	0.64	0.77	0.67	0.75
Sensitivity	0.46	0.70	0.61	0.68	0.61	0.69	0.66	0.66
Specificity (classified)*	0.74	0.84	0.67	0.82	0.74	0.92	0.80	0.94
Specificity	0.74	0.81	0.62	0.74	0.65	0.77	0.68	0.76
PCC (classified)*	0.61	0.78	0.65	0.77	0.69	0.85	0.74	0.84
PCC	0.61	0.76	0.61	0.71	0.63	0.74	0.67	0.71
AUC	0.61	0.78	0.65	0.75	0.66	0.74	0.67	0.72
PPV	0.67	0.82	0.66	0.80	0.72	0.91	0.78	0.92
NPV	0.59	0.75	0.64	0.74	0.66	0.80	0.70	0.79
Proportion Unclassified	0	0.02	0.05	0.07	0.08	0.13	0.09	0.15
Mean Classification Time	2.77	9.21	5.15	8.76	6.15	10.14	7.07	9.85

Note: These results were averaged across the five folds

[a] Model 1: Any JA-related visit

[b] Model 2: Any JA-related visit + Number of specialist visits

[c] Model 3: Any JA-related visit + Number of specialist visits + Hospitalization

[d] Model 4: Any JA-related visit + Number of specialist visits + Hospitalization + Number of general practitioner visits

*Measure is calculated without including unclassified individuals in the denominator

PCC: probability of correct classification

AUC: area under the receiver operating characteristic curve

PPV: positive predictive value

NPV: negative predictive value

For the random effects approach, the following comparisons amongst the four models can be made. Sensitivity of the classified data ranged from 0.71 (model 1) to 0.77 (model 3), sensitivity ranged from 0.66 (model 4) to 0.70 (model 1), specificity of the classified data ranged from 0.82 (model 2) to 0.94 (model 4), specificity ranged from 0.74 (model 2) to 0.81 (model 1), PCC of the classified data ranged from 0.77 (model 2) to 0.85 (model 3), PCC ranged from 0.71 (models 2 and 4) to 0.76 (model 1), AUC ranged from 0.72 (model 4) to 0.78 (model 1), PPV ranged from 0.80 (model 2) to 0.92 (model 4), and NPV ranged from 0.74 (model 2) to 0.80 (model 3). The proportion of unclassified individuals at the end of the time periods for models 1 to 4 were 0.02, 0.07, 0.13, and 0.15, respectively. Mean classification time was similar across the models, ranging from 8.76 years (model 2) to 10.14 years (model 4).

Model 1 with the random effects prediction approach was chosen as the best dynamic LoDA model, as it had the highest AUC amongst all the models at 0.78. This model was used for the remaining analyses.

Table 4.4 provides a summary of the classification accuracy for dynamic LoDA using model 1 by various prior probabilities of being a JA case (i.e., 0.30, 0.40, 0.50, 0.60, and 0.70). The marginal and random effects prediction approaches are shown. In general, the classification accuracy was similar across the prior group probabilities for both the marginal and random effects prediction approaches. However, the classification cut-offs changed to account for the different prior group probabilities. For simplicity, naïve prior group probabilities (i.e., prior group probabilities of 0.50 for both cases and non-cases) were used for the remaining analyses, as changing the prior group probabilities had very little impact on classification accuracy.

Table 4.4: Summary of the classification accuracy for dynamic LoDA by the prior probability of being a JA case (0.30, 0.40, 0.50, 0.60, 0.70) using model 1 (any JA-related visit) with the marginal and random effects prediction approaches

	Marginal					**Random Effects**				
	Prior Probability of JA Case					**Prior Probability of JA Case**				
	0.30	**0.40**	**0.50**	**0.60**	**0.70**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**
Cut-off	0.25	0.35	0.44	0.54	0.65	0.19	0.26	0.35	0.43	0.54
Sensitivity (classified)*	0.47	0.46	0.46	0.46	0.48	0.69	0.70	0.71	0.70	0.70
Sensitivity	0.47	0.46	0.46	0.46	0.48	0.69	0.70	0.70	0.70	0.70
Specificity (classified)*	0.68	0.74	0.74	0.74	0.69	0.85	0.84	0.84	0.82	0.81
Specificity	0.68	0.74	0.74	0.74	0.69	0.82	0.80	0.81	0.80	0.79
PCC (classified)*	0.57	0.61	0.61	0.61	0.59	0.77	0.77	0.78	0.76	0.76
PCC	0.57	0.61	0.61	0.61	0.59	0.76	0.76	0.76	0.75	0.75
AUC	0.60	0.61	0.61	0.61	0.61	0.78	0.78	0.78	0.78	0.78
PPV	0.58	0.67	0.67	0.67	0.59	0.82	0.81	0.82	0.79	0.78
NPV	0.58	0.59	0.59	0.59	0.58	0.74	0.75	0.75	0.75	0.74
Proportion Unclassified	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.01	0.01
Mean Classification Time	2.69	2.77	2.77	2.77	2.70	9.25	9.25	9.21	9.39	9.46

Note: These results are averaged across the five folds

*Measure is calculated without including unclassified individuals in the denominator

PCC: probability of correct classification

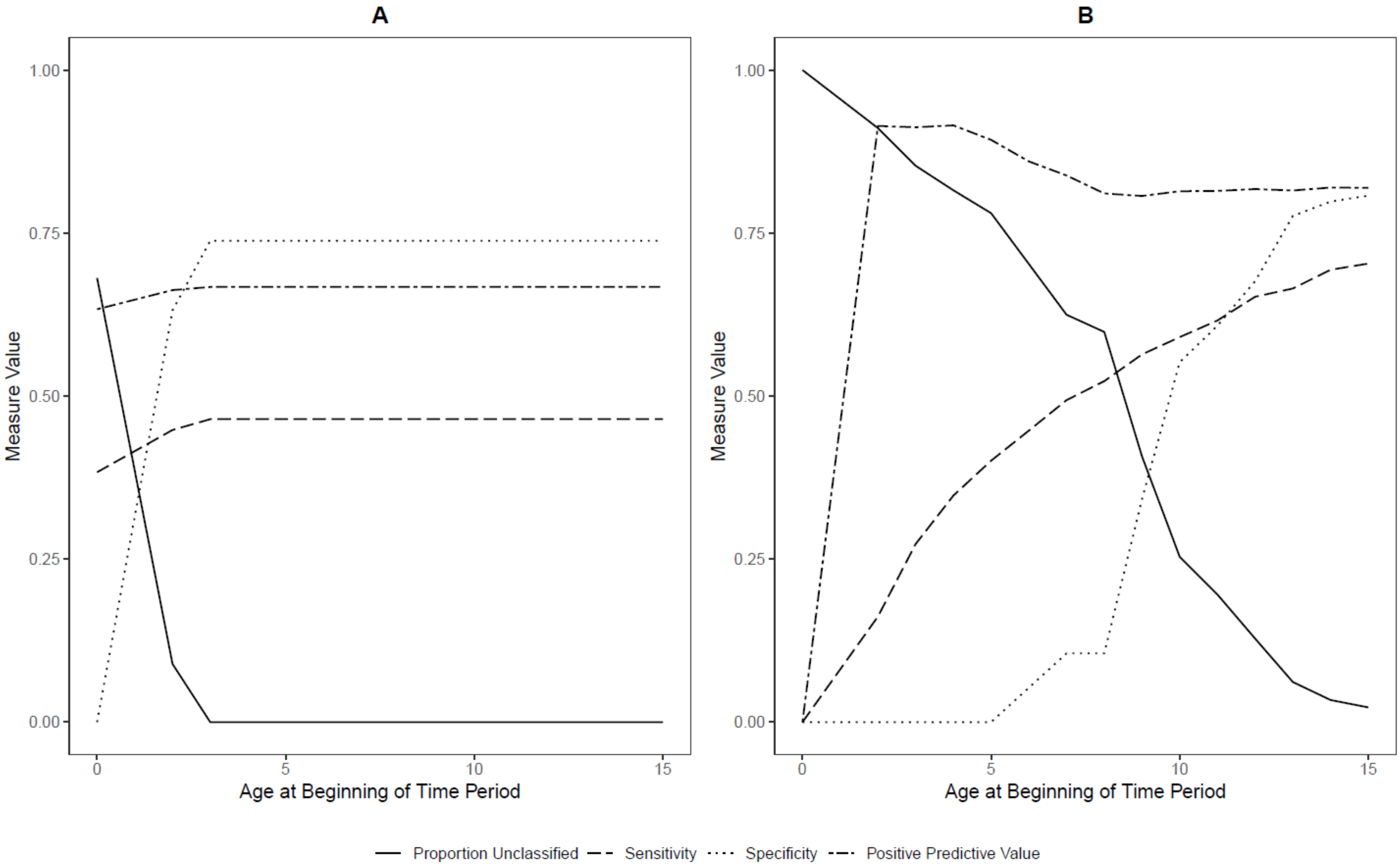
AUC: area under the receiver operating characteristic curve

PPV: positive predictive value

NPV: negative predictive value

Figure 4.3 illustrates how the proportion of individuals unclassified, sensitivity, specificity, and PPV changes over the time periods for the dynamic LoDA using model 1 with naïve prior probabilities and the marginal (Figure 4.3A) and random effects (Figure 4.3B) prediction approaches. For the marginal approach, 68.1% of individuals were left unclassified after the first time period (0 and 1 years of age), but dropped to only 8.9% of individuals after the second time period (2 years of age). By the end of the third time period (3 years of age), everyone was classified. Accordingly, sensitivity, specificity, PPV increased during the first three time periods and then stayed the same, as there were no new individuals that needed to be classified. For the random effects approach, all individuals were left unclassified after the first time period (0 and 1 years of age). The proportion of unclassified individuals then steadily decreased across all time periods where it reached 0.02 after the 15th time period (15 years of age). Sensitivity increased across all the time periods until it reached 0.70 after the last time period, whereas specificity remained at 0 until after the fifth time period (5 years of age) where it began to increase across the time periods until it reached 0.81 after the last time period. PPV jumped to 0.91 after the second time period and slightly decreased across the remaining time periods to 0.82.

Figure 4.3: Proportion of unclassified individuals, sensitivity, specificity, and PPV over time periods for dynamic LoDA using model 1 (any JA-related visit) with naïve prior probabilities and the A) marginal prediction approach and B) random effects approach



4.4 Static Model-Based and Deterministic Case Definitions

Table 4.5 summarizes the classification accuracy of the static LoDA models and deterministic case definition. The static LoDA model that used data from all 15 time periods (i.e., birth to 16th birthday) performed the best on all classification accuracy measures compared to the static LoDA models that used fewer periods of data, except for specificity, which was highest for the static model that used 5 time periods of data (i.e., birth to 6th birthday) at 0.91.

The deterministic JA case definition that used data from all 15 time periods (i.e., birth to 16th birthday) had high accuracy with a sensitivity, PCC, and NPV of 0.91 and a specificity and PPV of 0.92. When the deterministic case definition was applied only to data between the 14th to 16th birthdays, the sensitivity decreased by 37.4% to 0.57. The PCC and NPV also decreased, but slight increases in specificity and PPV were achieved.

When data from all 15 time periods were used (i.e., birth to 16th birthday), the deterministic case definition only slightly outperformed the static LoDA model. When only the data between the 14th and 16th birthdays were used, the static LoDA model was more sensitive with a 26.3% higher sensitivity than the deterministic case definition. However, the deterministic case definition was more specific.

Table 4.5: Summary of classification accuracy for dynamic LoDA compared to static LoDA and the deterministic JA case definition

	Dynamic LoDA	**Static LoDA**				**Deterministic**	
		Data Used for Classification				**Data Used for Classification**	
	Model 1	**Birth to 6th birthday**	**Birth to 11th birthday**	**Birth to 16th birthday**	**14th to 16th birthday**	**Birth to 16th birthday**	**14th to 16th birthday**
Cut-off	0.35	0.12	0.27	0.34	0.41	N/A	N/A
Sensitivity	0.70	0.44	0.61	0.89	0.72	0.91	0.57
Specificity	0.81	0.91	0.87	0.89	0.87	0.92	0.96
PCC	0.76	0.68	0.74	0.89	0.80	0.91	0.78
AUC	0.78	0.69	0.77	0.95	0.81	N/A	N/A
PPV	0.82	0.83	0.82	0.88	0.84	0.92	0.94
NPV	0.75	0.63	0.70	0.90	0.77	0.91	0.71
Proportion Unclassified	0.02	N/A	N/A	N/A	N/A	N/A	N/A

Note: The dynamic and static LoDA results presented were averaged over the 5 folds and used model 1 (any JA-related visit) with the random effects approach and naïve prior probabilities

PCC: probability of correct classification

AUC: area under the receiver operating characteristic curve

PPV: positive predictive value

NPV: negative predictive value

The dynamic LoDA model did not outperform the static LoDA model or the deterministic case definition that used all 15 periods of longitudinal data on any of the classification accuracy measures. The dynamic model was much less sensitive to detecting true JA cases with a sensitivity (0.70) that was 21.3% and 23.1% lower than the static model (0.89) and deterministic case definition (0.91), respectively. The dynamic model did, however, achieve higher sensitivity and NPV than the static models that used 5 time periods (i.e., birth to 6th birthday) and 10 time periods (i.e., birth to 11th birthday) of data. The dynamic model also achieved comparable PPV to the static models, but was still less specific than the static models.

4.5 Sex-Specific Classification

Table 4.6 summarizes the classification accuracy of the sex-specific dynamic and static LoDA models. For males, compared to the overall dynamic LoDA model, the sensitivity and NPV of the dynamic LoDA model were greater by 11.4 % and 22.7%, respectively. However, the specificity of the male dynamic model decreased by 16.0%. PPV remained the same, but the proportion unclassified increased by 600% (n=279).

For females, the dynamic LoDA model outperformed the dynamic model for both sexes on all classification accuracy measures. In particular, sensitivity and PPV increased by 8.6% and 7.3%, respectively. The proportion of unclassified individuals for the overall dynamic model (0.02) and female dynamic model (0.03) were similar. However, in terms of the static LoDA models, those developed for males and females had similar classification accuracy to the overall static LoDA model (Table 4.6).

Table 4.6: Summary of the classification accuracy for overall and sex-specific dynamic and static LoDA

	Dynamic LoDA			**Static LoDA**								
	Model 1			**Birth to 11th birthday**			**Birth to 16th birthday**			**14th to 16th birthday**		
	Overall	**Male**	**Female**	**Overall**	**Male**	**Female**	**Overall**	**Male**	**Female**	**Overall**	**Male**	**Female**
Cut-off	0.35	0.40	0.31	0.27	0.30	0.15	0.34	0.37	0.33	0.41	0.50	0.37
Sensitivity	0.70	0.78	0.76	0.61	0.58	0.63	0.89	0.87	0.90	0.72	0.69	0.73
Specificity	0.81	0.68	0.84	0.87	0.82	0.87	0.89	0.89	0.91	0.87	0.91	0.85
PCC	0.76	0.73	0.80	0.74	0.71	0.76	0.89	0.88	0.90	0.80	0.80	0.79
AUC	0.78	0.72	0.84	0.77	0.72	0.78	0.95	0.94	0.95	0.81	0.80	0.80
PPV	0.82	0.82	0.88	0.82	0.77	0.84	0.88	0.87	0.90	0.84	0.86	0.82
NPV	0.75	0.92	0.78	0.70	0.68	0.72	0.90	0.89	0.90	0.77	0.76	0.78
Proportion Unclassified	0.02	0.14	0.03	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Note: The dynamic and static LoDA results presented were averaged over the 5 folds and used model 1 (any JA-related visit) with the random effects approach and naïve prior probabilities

PCC: probability of correct classification

AUC: area under the receiver operating characteristic curve

PPV: positive predictive value

NPV: negative predictive value

CHAPTER 5 – DISCUSSION AND CONCLUSIONS

5.1 Summary

Dynamic classification, specifically dynamic LoDA, was adapted for use with retrospective longitudinal administrative health data to construct chronic disease case definitions. These methods were applied to Manitoba's administrative health data linked to the Pediatric Rheumatology Clinical Database to identify cases of JA. Classification accuracy of the best dynamic LoDA model was compared to a validated deterministic case definition and static LoDA models.

Four combinations of longitudinal disease markers (i.e., any JA-related visit, number of general practitioner visits, number of specialist visits, and hospitalization) were modelled separately for JA cases and non-cases using MGLMMs. Fixed effect covariates (i.e., sex, region of residence, and age) and random intercepts were included in the models to describe the evolution of the longitudinal disease markers over time. Differences in the values of the parameter estimates between JA cases and non-cases were most apparent for the disease marker of any JA-related visit.

Dynamic LoDA was applied to the data using each of the four models. Classifications were produced using three prediction approaches: marginal, conditional, and random effects. For each dynamic LoDA model, the random effects prediction approach produced higher classification accuracy than the marginal prediction approach. However, increased accuracy came at a cost of a higher proportion of unclassified individuals, as well as increased mean classification time.

For the marginal prediction approach, model 4 (i.e., all longitudinal disease markers) outperformed the other models on all classification accuracy measures, except for specificity.

The proportion of unclassified individuals and mean classification time increased incrementally from models 1 to 4.

For the random effects prediction approach, no dynamic LoDA model performed the best on all classification accuracy measures. Sensitivity and specificity calculated for the classified data, as well as PPV and NPV were higher for models 3 (i.e., any JA-related visit, number of specialist visits, and hospitalization) and 4 (i.e., all longitudinal disease markers), which left a higher proportion of individuals unclassified after all data were used. On the other hand, sensitivity and specificity calculated for all the data was highest for model 1 (i.e., any JA-related visit), which had the lowest proportion of unclassified individuals. Although the proportion of unclassified individuals increased as longitudinal disease markers were added to the models, the mean classification time remained similar.

Based on the AUC, model 1 (i.e., any JA-related visit) using the random effects prediction approach and naïve prior probabilities was the best dynamic LoDA model. Four additional hypothetical prior group probabilities were tested, but changing these probabilities had very little impact on classification accuracy.

Static LoDA models, using model 1 (i.e., any JA-related visit) with the random effects approach, were applied to the data to classify individuals as JA cases or non-cases using their entire longitudinal history after specified periods of time. The static LoDA model that used data from all 15 time periods (i.e., birth to 16th birthday) performed the best on most classification accuracy measures compared to the static LoDA models that used fewer periods of data. A validated deterministic JA case definition (Shiff et al., 2017) was also applied to the data. When all 15 time periods of data were used (i.e., birth to 16th birthday), the deterministic case definition slightly outperformed the static LoDA model.

The dynamic LoDA model did not outperform the static LoDA model or the deterministic case definition that used all 15 periods of data. In particular, the sensitivity of the dynamic model was much lower than both the static model and deterministic case definition. However, when fewer periods of data were used to construct the static LoDA models, the dynamic LoDA model achieved higher sensitivity and similar PPV.

Sex-specific dynamic models were produced by fitting the models separately for males and females. The model for males had greater sensitivity and NPV, but lower specificity than the overall dynamic model. The proportion of unclassified individuals was much larger for the male dynamic model than for the dynamic model for both sexes. The dynamic model for females outperformed the dynamic model for both sexes on all classification accuracy measures and the proportion of unclassified individuals was similar.

5.2 Discussion

The dynamic LoDA models developed in this study for JA case ascertainment did not outperform the static LoDA model or deterministic case definition (Shiff et al., 2017) applied to all time periods of data in terms of classification accuracy. This is not consistent with a previous application of dynamic classification that concluded that the dynamic classification approach had similar classification accuracy to the LoDA models that used all collected data (Hughes, Komárek, et al., 2018).

Although dynamic classification did not outperform the conventional deterministic case definition (Shiff et al., 2017) in this application to JA, there may be other chronic diseases where the dynamic classification approach is beneficial for case ascertainment in administrative health data. In particular, chronic diseases that have episodic symptoms and require many contacts with the healthcare system over time to be ascertained as a disease case in administrative health data, such as inflammatory bowel disease (Bernstein et al., 1999), may benefit from dynamic classification.

In this study, dynamic classification was applied to a pediatric population and a chronic condition that had well-defined age limits for diagnosis (i.e., birth to 16th birthday; Petty et al., 2004). These age limits defined the period of time for which the children were followed for classification. This will not be the case for all chronic conditions, especially chronic conditions in adult populations where it may not make sense or may not be possible to use an individual's complete longitudinal health history from birth. In these instances, the time period for which the population will be followed must be determined. One approach is to use a specified period of time after a defined index date, such as a treatment, procedure, or admission date (Duan et al., 2019; Hughes, Bonnett, et al., 2018; Lix et al., 2008; Van Walraven et al., 2010). Specified calendar or fiscal years or an age range of interest could also be used to define the period of follow-up (Choi et al., 2017). The time period chosen for follow-up will depend on the objectives of the study, the chronic disease and population of interest, as well as the data that are available.

In addition, the timing of classification updates needs to be explored. In this study, a standardized approach was adopted that could be applied to other chronic diseases; the first classification was conducted at the second birthday using the data accrued from birth, and then the classifications were updated annually until the 16th birthday. However, the classifications could have be updated monthly, biannually, or biennially. Implementing a fixed time period approach may result in challenges. For example, if the time periods are too short, there may not be sufficient data within the period to fit the MGLMMs. Sparse data may result inestimable model parameters and overfitting of the data. In this case, the length of the periods would need to be increased or the model(s) would need to be simplified. On the other hand, if the time periods are too long, the follow-up may not be sufficiently long for the classifications to be updated dynamically. Using the irregular time period definition, adopted in this study (i.e., the first time period was longer than the subsequent time periods), may help to address these challenges.

An individualized updating schedule, similar to that used in prospective applications of dynamic classification (Hughes, Bonnett, et al., 2018; Hughes et al., 2017; Hughes, Komárek, et al., 2018), could also be used. This would involve updating an individual's classifications each time a new health care visit is recorded in the administrative health data. Researchers must carefully consider the optimal updating approach for their application based on the features of the data available to them and the characteristics of their outcome of interest.

In regard to the development of the MGLMMs, several modelling decisions need to be made. The longitudinal disease markers of interest and their corresponding distributions must be chosen. If the MGLMMs are fit with the mixAK package, only three distribution options (i.e., Gaussian, Bernoulli, and Poisson) are currently available (Komárek & Komárková, 2014). The combinations of the longitudinal disease markers jointly modelled must also be selected. Different combinations of longitudinal disease markers can have varying classification accuracy and considering multiple longitudinal disease markers may not improve classification accuracy (Hughes, Komárek, et al., 2018). In this study, the selection of longitudinal disease markers (i.e. models 1 to 4) was made based on the descriptive analyses of the trends of the longitudinal disease markers (Figure 4.2). The longitudinal disease markers were sequentially added to the models based on visual assessment of the linear trends for JA cases and non-cases.

In this study, sex, age at the beginning of the time period, and region of residence at birth were chosen as model covariates. If covariates had not been included in the models, variation in the longitudinal disease markers would only be explained by the random intercepts. The covariates were chosen *a priori* because of their association with a JA diagnosis and healthcare utilization (Brownell et al., 2002; Shiff et al., 2014, 2019).

In addition, the types of random effects included in the models, as well their corresponding distribution must be chosen. Both random intercepts and random slopes can be included in the MGLMMs. Due to the smaller sample size in this study, only random intercepts were included in the models to prevent overfitting. The MGLMM adopts a normal mixture distribution for the random effects to robustify the model against misspecification of the random effects distribution. The number of mixture components for the random effects distribution must be chosen by the investigator. The optimal number of mixture components can be determined by assessing model fit using the PED, where lower values of the PED indicate better model fit (Plummer, 2008). Classification accuracy can vary with the choice of the number of mixture components (Hughes, Komárek, et al., 2018).

5.3 Implications

The results of this study suggest that dynamic classification can produce accurate chronic disease case definitions using longitudinal information from administrative health data. Currently, the deterministic approach is the most widely-used approach to construct case definitions for population-based chronic disease research and surveillance (Lix et al., 2006; Quan et al., 2009; Shiff et al., 2017). Benefits of using a model-based approach for case ascertainment have also been presented, but most aggregate longitudinal information (Cooke et al., 2011; Lix et al., 2008; Van Walraven et al., 2010). This study has shown that there can be value in using longitudinal information in a dynamic manner to create case definitions. Instead of viewing case definition development as “one approach fits all”, researchers should carefully examine the characteristics of their disease of interest, as well as the data available to determine which approach, deterministic, model-based, or dynamic, may result in the greatest classification accuracy.

In the context of population-based chronic disease surveillance, researchers need to determine the implications of classifying an individual as a disease case or non-case if there is insufficient evidence to do so. Dynamic classification allows individuals to remain unclassified if specified criteria indicate that more information is needed to make a potentially more accurate decision. Allowing individuals to remain unclassified may allow for more accurate and informative population chronic disease patterns to emerge.

Many mature administrative health data repositories exist in Canada. These include the MCHP in Manitoba, Population Data BC in British Columbia, and the Institute for Clinical Evaluative Sciences (IC/ES) in Ontario. Each of these repositories holds population-based longitudinal administrative health data from 1970 onwards, 1985 onwards, and 1991 onwards, respectively (Institute for Clinical Evaluative Sciences, 2019; Manitoba Centre for Health Policy, 2019; Population Data BC, 2019). More efficient use of these data repositories is possible to create cohorts that can be used with longitudinal statistical methods, such as dynamic classification, which have the potential to produce more accurate results.

This study also revealed differences in how dynamic classification is applied to retrospective administrative health data compared to prospective clinical data. First, when to update classifications is not as straightforward when using retrospective administrative health data compared to prospective clinical data. For prospective clinical data, classifications are updated each time an individual has a follow-up appointment with their physician (Hughes, Bonnett, et al., 2018; Hughes et al., 2017; Hughes, Komárek, et al., 2018). When using retrospective administrative health data, the investigator needs to define when the classifications will be updated, whether that be based on time periods or visits recorded in the data. Second, the types of variables defined from administrative and clinical data differ. Symptoms, laboratory results, and risk factors can be extracted from clinical data, but this information is not found in administrative health data. Dynamic models using retrospective administrative health data must rely on variables defined from healthcare utilization and recorded diagnosis codes. As shown in this study, variables defined from administrative health data tend to be binary or count variables, which can result in challenges related to sparse data.

5.4 Strengths and Limitations

This study has both limitations and strengths. With respect to the former, first, once an individual was classified as a JA case or non-case, their status was not revisited in additional data periods. For example, if an individual had strong evidence of not having JA earlier in their life and was ultimately classified as a non-case, their status was not revisited when that individual was older. This is a limitation of case ascertainment, especially for rheumatic diseases, which can evolve over time for some patients. Applying the static LoDA model to the longitudinal data at different points in time from birth displayed that classification accuracy, notably sensitivity, AUC, and NPV, increased as more years of data after birth were used. In addition, the static LoDA model applied between the 14th and 16th birthdays had higher classification accuracy than the static LoDA models applied earlier in life, implying that the classifications made by the dynamic LoDA models early in life may be less accurate.

Second, only one updating approach was used to make classifications. Applying a different updating approach to the study data, such as updating the results each time a new visit is recorded in the administrative health data, could influence the modelling and classification results.

Third, the study cohort was a referral population defined from the Pediatric Rheumatology Clinical Database and may not represent the general population. Therefore, the estimates for specificity will not generalize to the general population.

The main strength of this study was the population-based administrative health data and clinical data used. Both inpatient and outpatient records were available for all members of the study cohort from birth until 16th birthday. In addition, during the time frame of this study, there was only one pediatric rheumatology centre serving Manitoba, meaning children with potential rheumatic disease diagnoses were not being referred elsewhere. Therefore, the Pediatric Rheumatology Clinical Database captures almost all pediatric rheumatic disease cases in the province and having the clinical diagnoses recorded by a single practitioner for the entire study period allowed for no inter-physician variability.

Furthermore, this study was based on a previous validation study that developed JA case definitions using the same clinical database (Shiff et al., 2017). The validation study's cohort and methods were well documented and provided guidance for this study. It also allowed for the dynamic models to be compared to a well-defined deterministic case definition.

5.5 Opportunities for Future Research

This research leads to many opportunities for future research. First, dynamic classification methods should be applied to other populations and chronic diseases, such as inflammatory bowel disease. Applying dynamic classification to other populations and chronic diseases will help provide guidance and recommendations for the contexts in which dynamic classification will work best. Second, the feasibility of using other updating schedules, such as every time a visit is recorded in the administrative health data, should be explored.

Third, different classification rules could be used. In this study an individual was classified once the cut-off fell outside of the 99% CrI for the probability of being a JA case. Hughes et al. (2017) presented a list of other potential classification rules, including an individual requiring two consecutive high probabilities to be classified as a disease case. Changing the rules for when an individual is classified could influence classification accuracy.

Fourth, different models could be used within the dynamic classification scheme. This study used LoDA to classify individuals based on their longitudinal history. However, other models that can be applied to longitudinal data, such as mixed effects logistic regression models or machine learning models, could be used with dynamic classification. In addition, using multistate models in the context of dynamic classification may allow for individuals to move between disease case and non-case groups over time (Hughes et al., 2017).

REFERENCES

Albert, P. S. (2012). A linear mixed model for predicting a binary event from longitudinal data under random effects misspecification. *Statistics in Medicine*, *31*(2), 143–154.

Bernstein, C. N., Blanchard, J. F., Rawsthorne, P., & Wajda, A. (1999). Epidemiology of Crohn's disease and ulcerative colitis in a central Canadian province: a population-based study. *American Journal of Epidemiology*, *149*(10), 916–924.

Bernstein, C. N., Wajda, A., Svenson, L. W., MacKenzie, A., Koehoorn, M., Jackson, M., … Blanchard, J. F. (2006). The epidemiology of inflammatory bowel disease in Canada: A population-based study. *American Journal of Gastroenterology*, *101*(7), 1559–1568.

Brant, L. J., Sheng, S. L., Morrell, C. H., Verbeke, G. N., Lesaffre, E., & Carter, H. B. (2003). Screening for prostate cancer by using random-effects models. *Journal of the Royal Statistical Society Series A*, *166*(1), 51–62.

Brant, L. J., Sheng, S. L., Morrell, C. H., & Zonderman, A. B. (2005). Data from a longitudinal study provided measurements of cognition to screen for Alzheimer's disease. *Journal of Clinical Epidemiology*, *58*, 701–707.

Brewer, E. J., Bass, J., Cassidy, J. T., Fink, C., Jacobs, J., Hanson, V., … Stillman, J. S. (1977). Current proposed revision of JRA criteria. *Arthritis and Rheumatism*, *20*(2 Suppl), 195–199.

Brownell, M., Kozyrskyj, A., Roos, N. P., Friesen, D., Mayer, T., Sullivan, K., & Brownell, M. D. (2002). Health service utilization by Manitoba children. *Canadian Journal of Public Health*, *93*(Supplement 2), S57–S62.

Cadarette, S. M., & Wong, L. (2015). An introduction to health care administrative data. *The Canadian Journal of Hospital Pharmacy*, *68*(3), 232–237.

Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, *24*(2), 361–370.

Cooke, C. R., Joo, M. J., Anderson, S. M., Lee, T. A., Udris, E. M., Johnson, E., & Au, D. H. (2011). The validity of using ICD-9 codes and pharmacy records to identify patients with chronic obstructive pulmonary disease. *BMC Health Services Research*, *11*(37), 1–10.

Dart, A. B., Martens, P. J., Sellers, E. A., Brownell, M. D., Rigatto, C., & Dean, H. J. (2011). Validation of a pediatric diabetes case definition using administrative health data in Manitoba, Canada. *Diabetes Care*, *34*(4), 898–903.

Duan, H., Sun, Z., Dong, W., & Huang, Z. (2019). Utilizing dynamic treatment information for MACE prediction of acute coronary syndrome. *BMC Medical Informatics and Decision Making*, *19*(5), 1–11.

Elliott, M. R., Sammel, M. D., & Faul, J. (2012). Associations between variability of risk factors and health outcomes in longitudinal studies. *Statistics in Medicine*, *31*(23), 2745–2756.

English, S. W., McIntyre, L., Fergusson, D., Turgeon, A., dos Santos, M. P., Lum, C., … van Walraven, C. (2016). Subarachnoid hemorrhage admissions retrospectively identified using a prediction model. *Neurology*, *87*, 1557–1564.

Fan, J., Arruda-Olson, A. M., Leibson, C. L., Smith, C., Liu, G., Bailey, K. R., & Kullo, I. J. (2013). Billing code algorithms to identify cases of peripheral artery disease from administrative data. *Journal of the American Medical Informatics Association*, *20*, e349–e354.

García-Fiñana, M., Hughes, D. M., Cheyne, C. P., Broadbent, D. M., Wang, A., Komárek, A., … Harding, S. P. (2018). Personalized risk-based screening for diabetic retinopathy: A multivariate approach versus the use of stratification rules. *Diabetes, Obesity and Metabolism*, 1–9.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. In *Statistical Science* (Vol. 7).

Harrold, L. R., Salman, C., Shoor, S., Curtis, J. R., Asgari, M. M., Gelfand, J. M., … Herrinton, L. J. (2013). Incidence and prevalence of juvenile idiopathic arthritis among children in a managed care population, 1996-2009. *The Journal of Rheumatology*, *40*(7), 1218–1225.

Horrocks, J., & van Den Heuvel, M. J. (2009). Prediction of pregnancy: A joint model for longitudinal and binary data. *Bayesian Analysis*, *4*(3), 523–538.

Hua, Y., Jeske, D. R., Paul, R., & James, B. (2010). Neutral zone classifiers using a decision-theoretic approach with application to DNA array analyses. *Journal of Agricultural, Biological, and Environmental Statistics*, *15*(4), 474–490.

Hughes, D. M., Bonnett, L. J., Czanner, G., Komárek, A., Marson, A. G., & García-Fiñana, M. (2018). Identification of patients who will not achieve seizure remission within 5 years on AEDs. *Neurology*, *91*(22), 2035–2044.

Hughes, D. M., El Saeiti, R., & García-Fiñana, M. (2018). A comparison of group prediction approaches in longitudinal discriminant analysis. *Biometrical Journal*, *60*, 307–322.

Hughes, D. M., Komárek, A., Bonnet, L. J., Czanner, G., & García-Fiñana, M. (2017). Dynamic classification using crediblie intervals in longitudinal discriminant analysis. *Statistics in Medicine*, *36*, 3858–3874.

Hughes, D. M., Komárek, A., Czanner, G., & García-Fiñana, M. (2018). Dynamic longitudinal discriminant analysis using multiple longitudinal markers of different types. *Statistical Methods in Medical Research*, *27*(7), 2060–2080.

Institute for Clinical Evaluative Sciences. (2019). ICES Data. Retrieved September 27, 2019, from https://www.ices.on.ca/Data-and-Privacy/ICES-data

Jeske, D. R., Linehan, J. A., Wilson, T. G., Kawachi, M. H., Wittig, K., Lamparska, K., … Smith, S. S. (2017). Two-stage classifiers that minimize PCA3 and the PSA proteolytic activity testing in the prediction of prostate cancer recurrence after radical prostatectomy. *The Canadian Journal of Urology*, *24*(6), 9089–9097.

Jeske, D. R., Liu, Z., Bent, E., & Borneman, J. (2007). Classification rules that include neutral zones and their application to microbial community profiling. *Communications in Statistics - Theory and Methods*, *36*(10), 1965–1980.

Jeske, D. R., & Smith, S. (2018). Maximizing the usefulness of statistical classifiers for two populations with illustrative applications. *Statistical Methods in Medical Research*, *27*(8), 2344–2358.

Jiang, B., Elliott, M. R., Sammel, M. D., & Wang, N. (2015). Joint modeling of cross-sectional health outcomes and longitudinal predictors via mixtures of means and variances. *Biometrics*, *71*(2), 487–497.

Jutte, D. P., Roos, L. L., & Brownell, M. D. (2011). Administrative record linkage as a tool for public health research. *Annual Review of Public Health*, *32*, 91–108.

Komárek, A., Hansen, B. E., Kuiper, E. M. M., van Buuren, H. R., & Lesaffre, E. (2010). Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Statistics in Medicine*, *29*, 3267–3283.

Komárek, A., & Komárková, L. (2013). Clustering for multivariate continuous and discrete longitudinal data. *The Annals of Applied Statistics*, *7*(1), 177–200.

Komárek, A., & Komárková, L. (2014). Capabilities of R Package mixAK for Clustering Based on Multivariate Continuous and Discrete Longitudinal Data. *Journal of Statistical Software*, *59*(12), 1–38.

Leslie, W. D., Lix, L. M., & Yogendran, M. S. (2011). Validation of a case definition for osteoporosis disease surveillance. *Osteoporosis International*, *22*(1), 37–46.

Lix, L. M., & Sajobi, T. T. (2010). Discriminant analysis for repeated measures data: A review. *Frontiers in Psychology*, *1*(146), 1–9.

Lix, L. M., Yogendran, M., Burchill, C., Metge, C., Mckeen, N., Moore, D., & Bond, R. (2006). *Defining and validating chronic diseases: an administrative data approach*. Winnipeg: Manitoba Centre for Health Policy.

Lix, L. M., Yogendran, M. S., Leslie, W. D., Shaw, S. Y., Baumgartner, R., Bowman, C., … James, R. C. (2008). Using multiple data features improved the validity of osteoporosis case ascertainment from administrative databases. *Journal of Clinical Epidemiology*, *61*(12), 1250–1260.

Manitoba Centre for Health Policy. (2019). Manitoba Population Data Repository Available Years of Data. Retrieved September 27, 2019, from http://umanitoba.ca/faculties/health_sciences/medicine/units/chs/departmental_units/mchp/resources/repository/datalist.html

Manuel, D. G., Rosella, L. C., & Stukel, T. A. (2010). Importance of accurately identifying chronic disease in studies using electronic health records. *British Medical Journal*, *341*(7770), 440–443.

Marshall, G., De La Cruz-Mesía, R., Quintana, F. A., & Barón, A. E. (2009). Discriminant analysis for longitudinal data with multiple continuous responses and possibly missing data. *Biometrics*, *65*(1), 69–80.

Maruyama, N., Takahashi, F., & Takeuchi, M. (2009). Prediction of an outcome using trajectories estimated from a linear mixed model. *Journal of Biopharmaceutical Statistics*, *19*(5), 779–790.

Mohd Din, S. H., Molas, M., Luime, J., & Lesaffre, E. (2014). Longitudinal profiles of bounded outcome scores as predictors for disease activity in rheumatoid arthritis patients: A joint modeling approach. *Journal of Applied Statistics*, *41*(8), 1627–1644.

Morrell, C. H., Brant, L. J., Sheng, S., & Metter, E. J. (2012). Screening for prostate cancer using multivariate mixed-effects models. *Journal of Applied Statistics*, *39*(6), 1151–1175.

O'Donnell, S., & Canadian Chronic Disease Surveillance System (CCDSS) Osteoporosis Working Group. (2013). Use of administrative data for national surveillance of osteoporosis and related fractures in Canada: Results from a feasibility study. *Archives of Osteoporosis*, *8*(143), 1–6.

Peng, M., Chen, G., Lix, L. M., McAlister, F. A., Tu, K., Campbell, N. R., … Hypertension Outcomes Surveillance Team. (2015). Refining hypertension surveillance to account for potentially misclassified cases. *PLoS ONE*, *10*(3), 1–11.

Petty, R. E., Laxer, R. M., Lindsley, C. B., & Wedderburn, L. (2016). *Textbook of pediatric rheumatology*. Philadelphia: Elsevier.

Petty, R. E., Southwood, T. R., Manners, P., Baum, J., Glass, D. N., Goldenberg, J., … International League of Associations for Rheumatology. (2004). International League of Associations for Rheumatology classification of juvenile idiopathic arthritis: second revision, Edmonton, 2001. *The Journal of Rheumatology*, *31*(2), 390–392.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, *9*(3), 523–539.

Population Data BC. (2019). Data available. Retrieved September 27, 2019, from https://www.popdata.bc.ca/data

Quan, H., Khan, N., Hemmelgarn, B. R., Tu, K., Chen, G., Campbell, N., … McAlister, F. A. (2009). Validation of a case definition to define hypertension using administrative data. *Hypertension*, *54*(6), 1423–1428.

Reddy, B. K., & Delen, D. (2018). Predicting hospital readmission for lupus patients: An RNN-LSTM-based deep-learning methodology. *Computers in Biology and Medicine, 101*, 199–209.

Robitaille, C., Dai, S., Waters, C., Loukine, L., Bancej, C., Quach, S., … Quan, H. (2012). Diagnosed hypertension in Canada: incidence, prevalence and associated mortality. *Canadian Medical Association Journal*, *184*(1), E49-56.

Shiff, N. J., Lix, L. M., Oen, K., Joseph, L., Duffy, C., Stringer, E., … Bernatsky, S. (2014). Chronic inflammatory arthritis prevalence estimates for children and adolescents in three Canadian provinces. *Rheumatology International*, *35*(2), 345–350.

Shiff, N. J., Oen, K., Kroeker, K., & Lix, L. M. (2019). Trends in population-based incidence and prevalence of juvenile idiopathic arthritis in Manitoba, Canada. *Arthritis Care & Research*, *71*(3), 413–418.

Shiff, N. J., Oen, K., Rabbani, R., & Lix, L. M. (2017). Validation of administrative case ascertainment algorithms for chronic childhood arthritis in Manitoba, Canada. *Rheumatology International*, *37*(9), 1575–1584.

Smith, M., Lix, L. M., Azimaee, M., E Enns, J., Orr, J., Hong, S., & Roos, L. R. (2017). Assessing the quality of administrative data for research: a framework from the Manitoba Centre for Health Policy. *Journal of the American Medical Informatics Association*, 1–6.

Stringer, E., & Bernatsky, S. (2015). Validity of juvenile idiopathic arthritis diagnoses using administrative health data. *Rheumatology International*, *35*(3), 575–579.

Van Walraven, C., Austin, P. C., Manuel, D., Knoll, G., Jennings, A., & Forster, A. J. (2010). The usefulness of administrative databases for identifying disease cohorts is increased with a multivariate model. *Journal of Clinical Epidemiology*, *63*(12), 1332–1341.

Virnig, B. A., & McBean, M. (2001). Administrative data for public health surveillance and planning. *Annual Review of Public Health*, *22*, 213–230.

Wang, T., Qiu, R. G., & Yu, M. (2018). Predictive modeling of the progression of Alzheimer's Disease with recurrent neural networks. *Scientific Reports*, *8*(9161), 1–12.

Wilchesky, M., Tamblyn, R. M., & Huang, A. (2004). Validation of diagnostic codes within medical services claims. *Journal of Clinical Epidemiology*, *57*(2), 131–141.

Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Informatics Association*, *25*(10), 1419–1428.

Yao, Z.-J., Bi, J., & Chen, Y.-X. (2018). Applying deep learning to individual and community health monitoring data: A survey. *International Journal of Automation and Computing*, *15*(6), 643–655.

Zeng, C., Ellis, J. L., Steiner, J. F., Shoup, J. A., McQuillan, D. B., & Bayliss, E. A. (2014). Assessment of morbidity over time in predicting health outcomes. *Medical Care*, *52*(3), S52–S59.

Zhang, X., Jeske, D. R., Li, J., & Wong, V. (2016). A sequential logistic regression classifier based on mixed effects with applications to longitudinal data. *Computational Statistics and Data Analysis*, *94*, 238–249.

APPENDIX A: LINEAR TRENDS OF LONGITUDINAL DISEASE MARKERS STRATIFIED BY COVARIATES

Figure A.1: Linear trends stratified by JA cases and non-cases for any JA-related visit for A) Males, B) Females, C) Winnipeg residents at birth, and D) Non-Winnipeg residents at birth

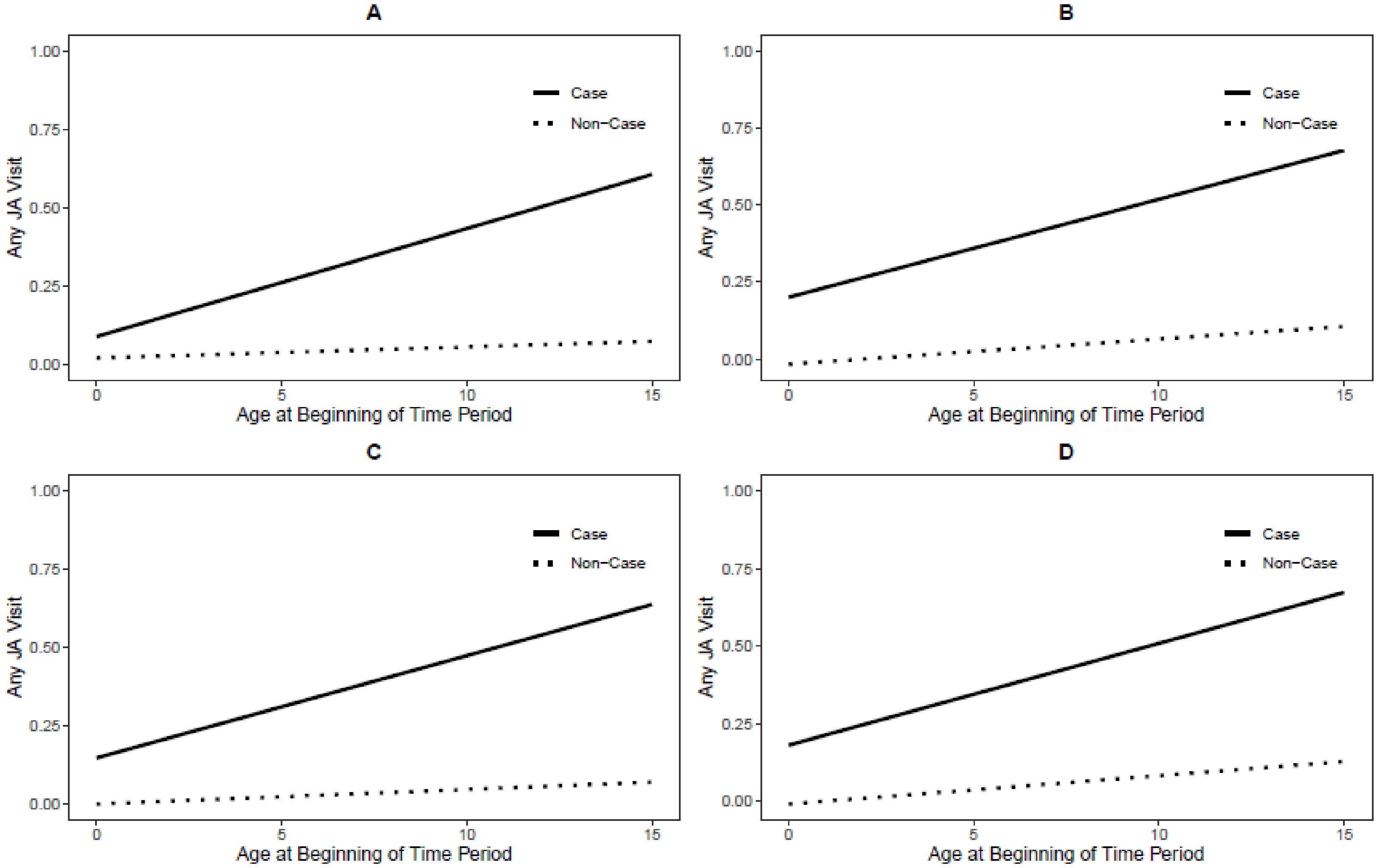


Figure A.2: Linear trends stratified by JA cases and non-cases for number of general practitioner visits for A) Males, B) Females, C) Winnipeg residents at birth, and D) Non-Winnipeg residents at birth

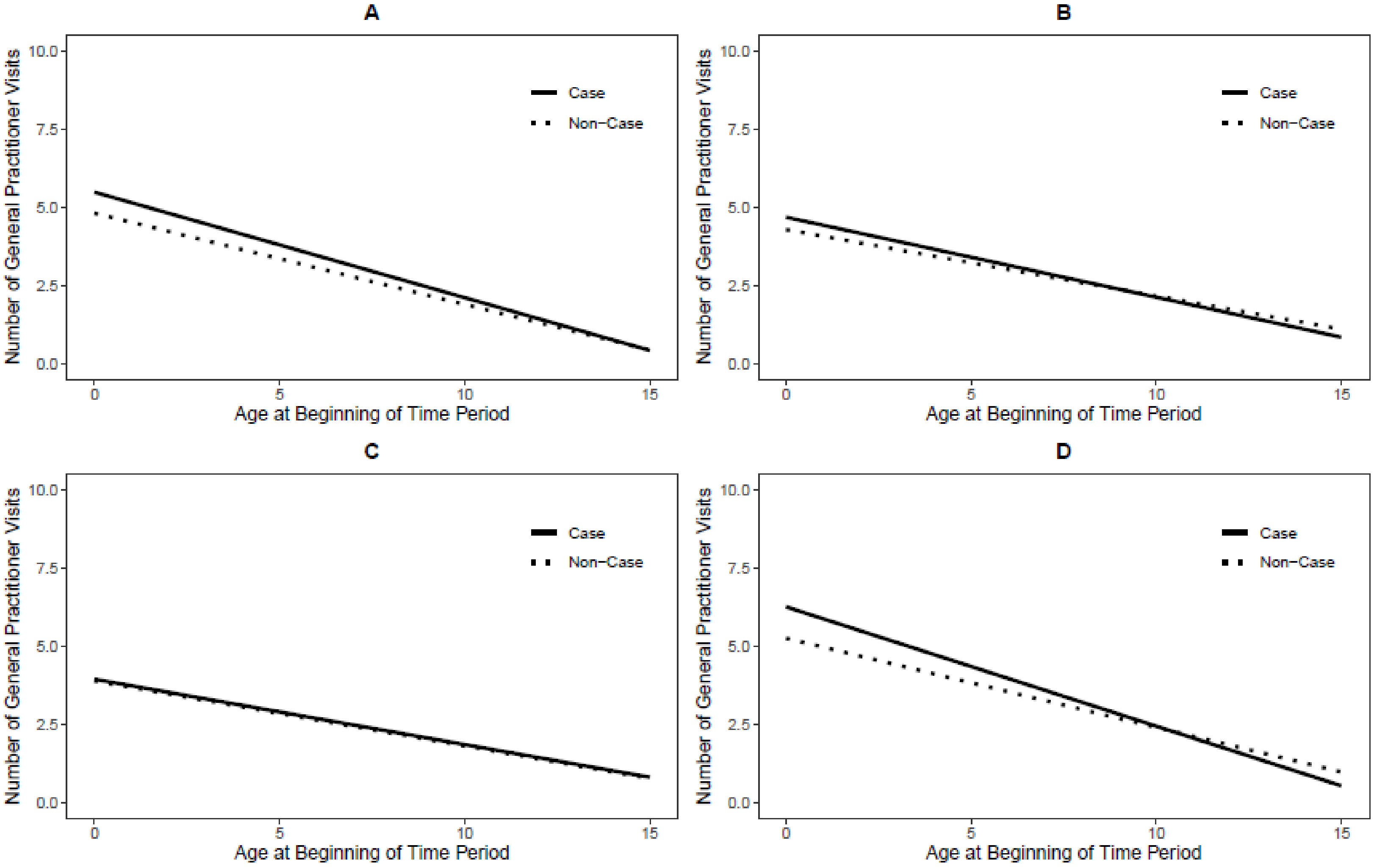


Figure A.3: Linear trends stratified by JA cases and non-cases for number of specialist visits for A) Males, B) Females, C) Winnipeg residents at birth, and D) Non-Winnipeg residents at birth

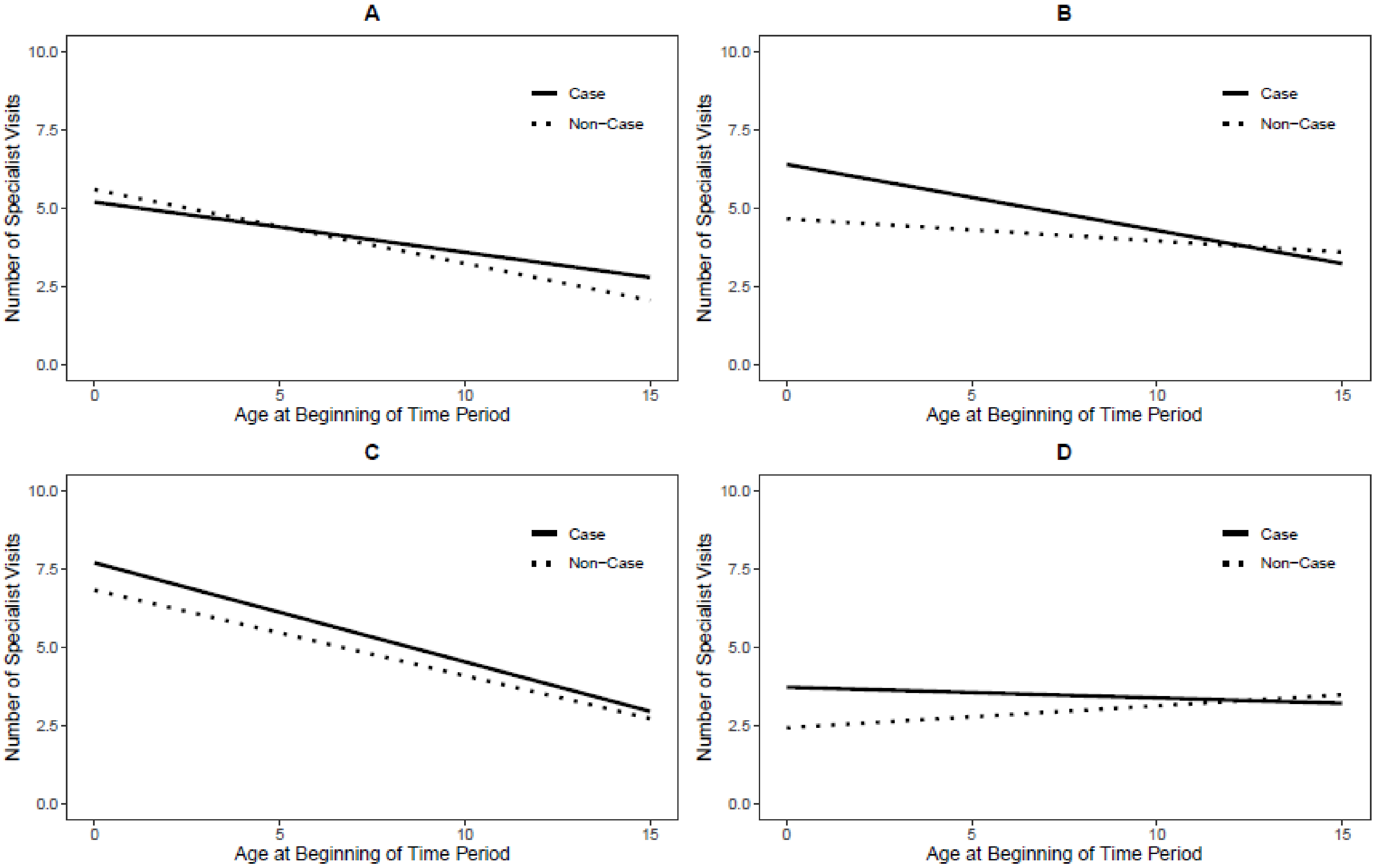
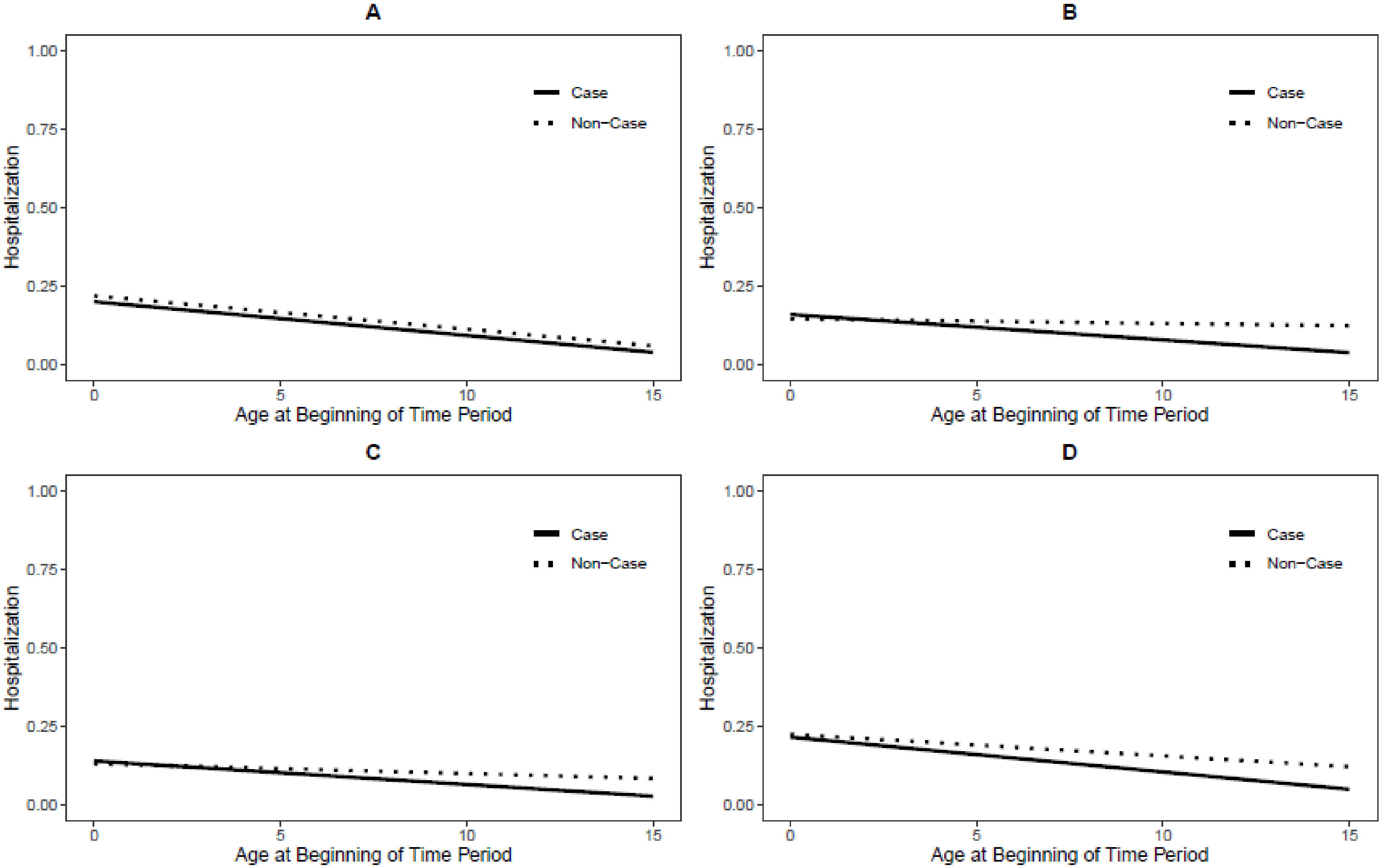


Figure A.4: Linear trends stratified by JA cases and non-cases for hospitalization for A) Males, B) Females, C) Winnipeg residents at birth, and D) Non-Winnipeg residents at birth



APPENDIX B: MODEL CONVERGENCE DIAGNOSTIC PLOTS

Figure B.1: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for any JA-related visit in model 1 (JA cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept

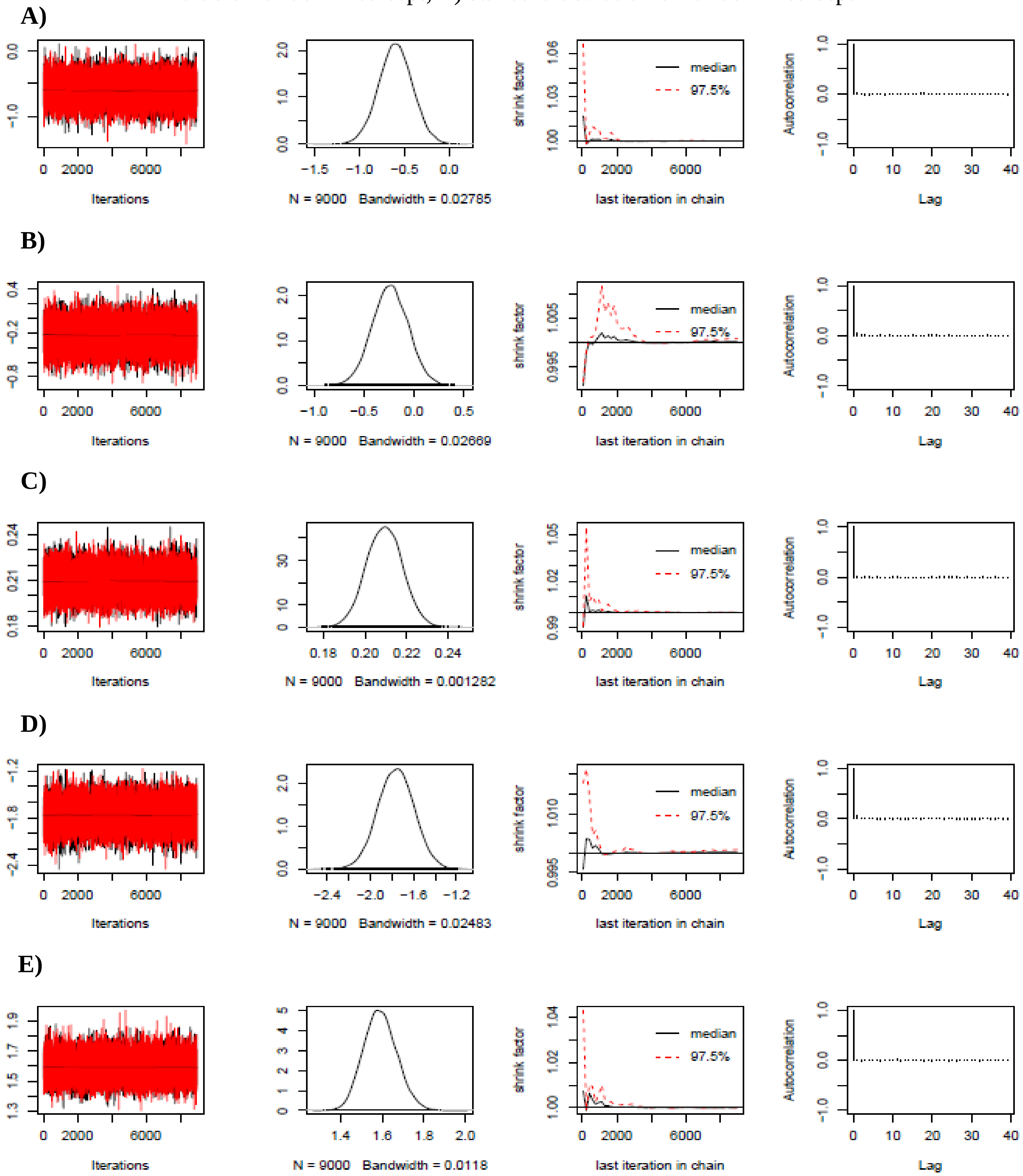


Figure B.2: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for any JA-related visit in model 1 (JA non-cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept

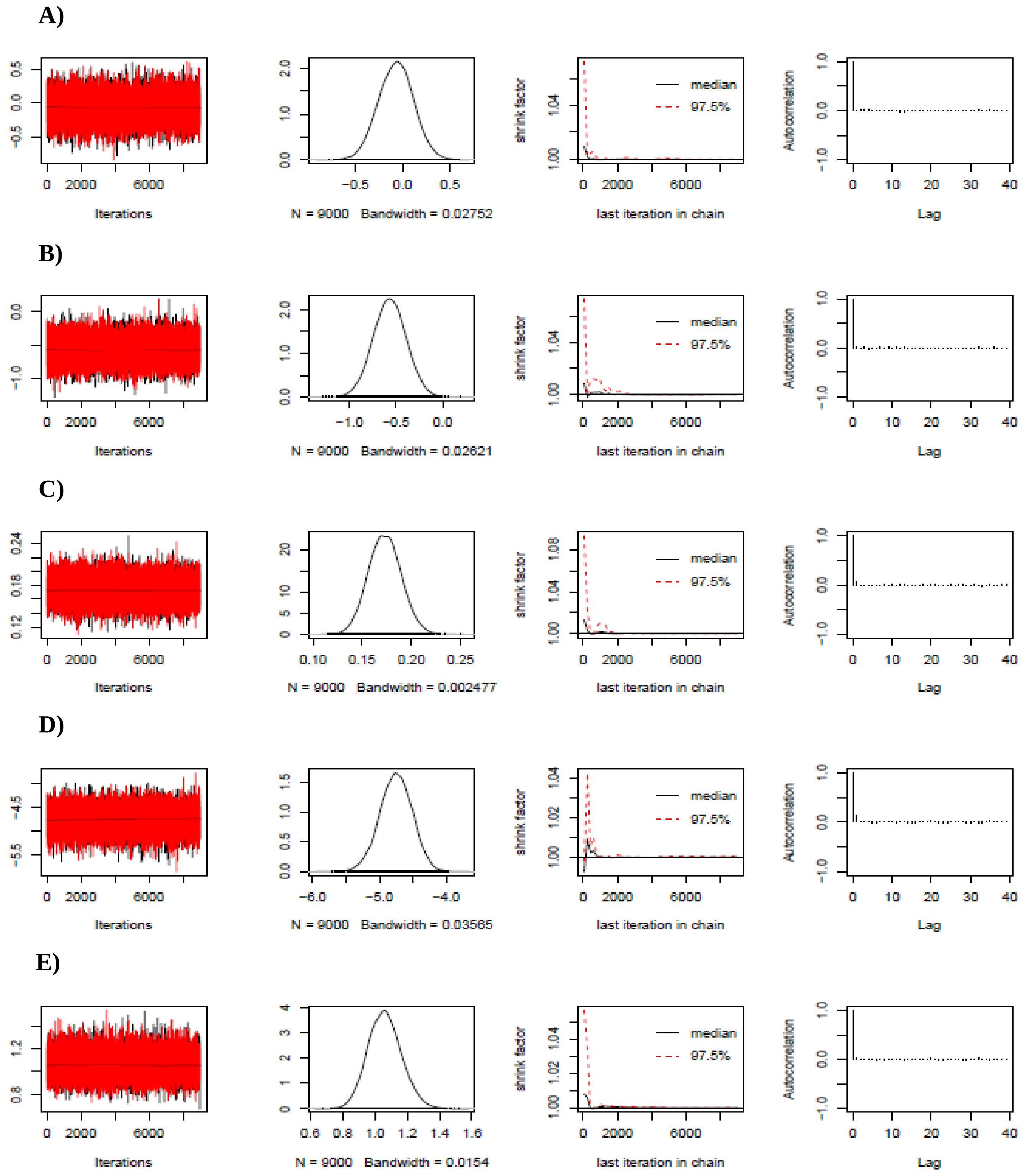


Figure B.3: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for any JA-related visit in model 2 (JA cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept

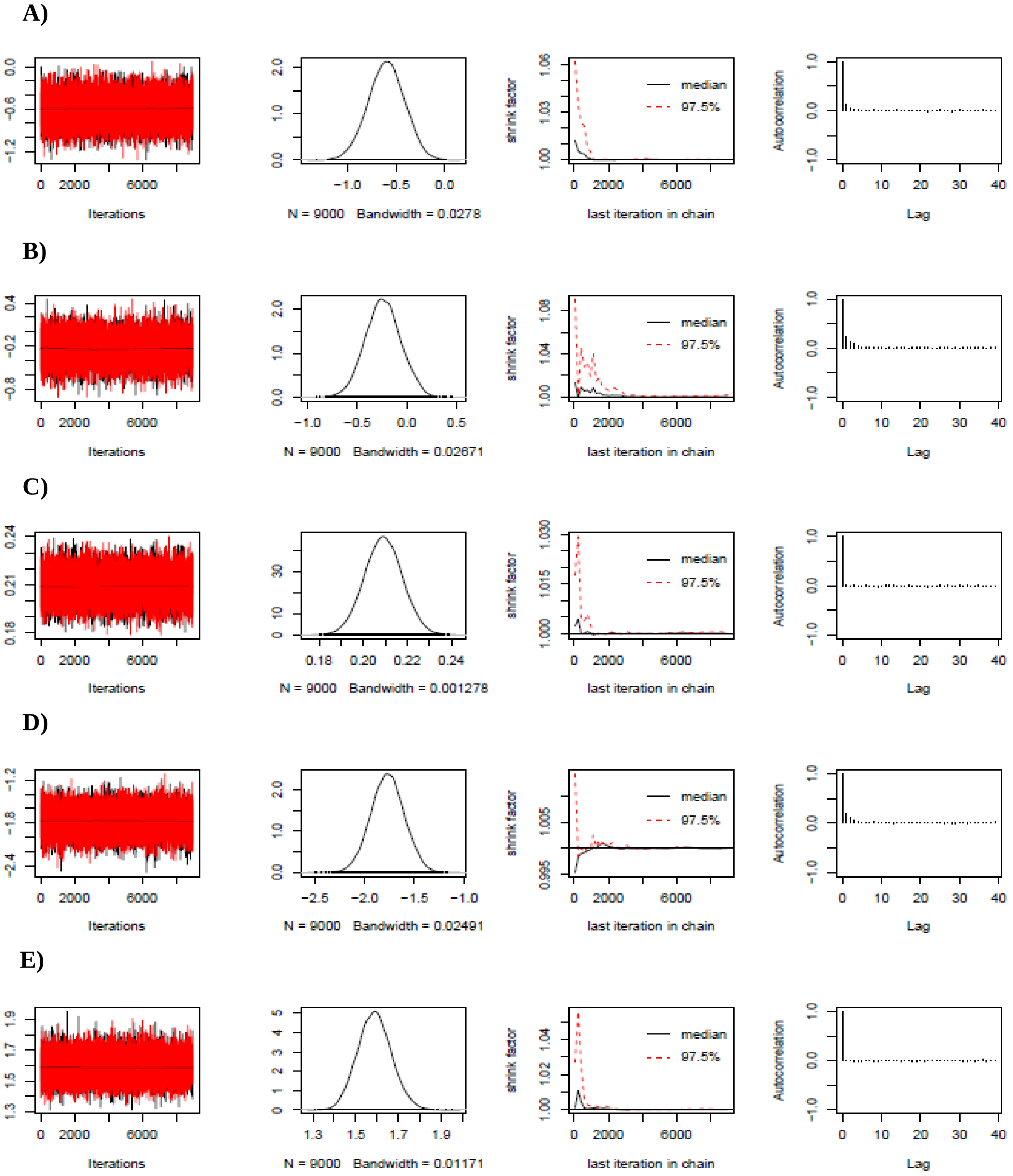


Figure B.4: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for any JA-related visit in model 2 (JA non-cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept

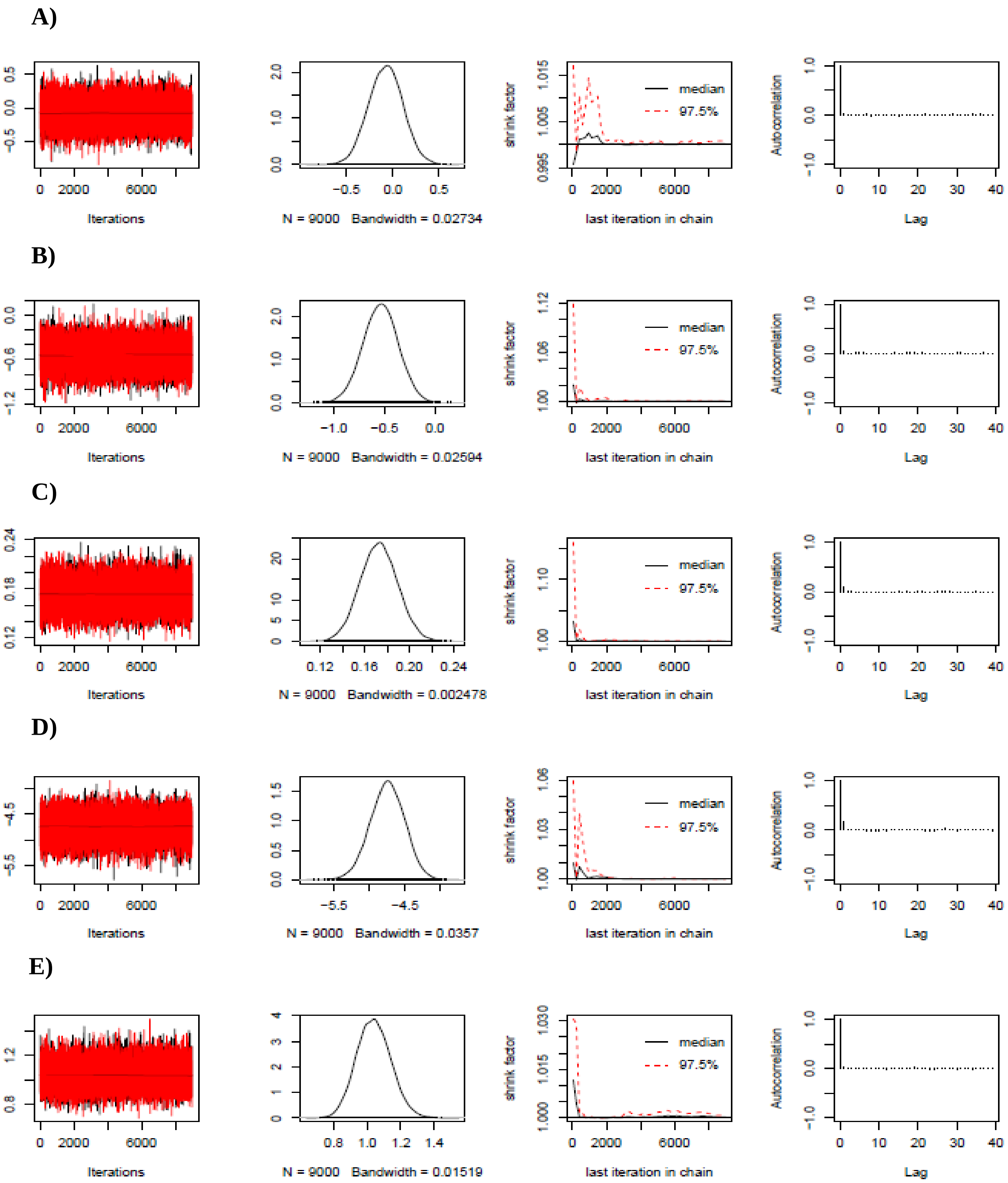


Figure B.5: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for number of specialist visits in model 2 (JA cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept

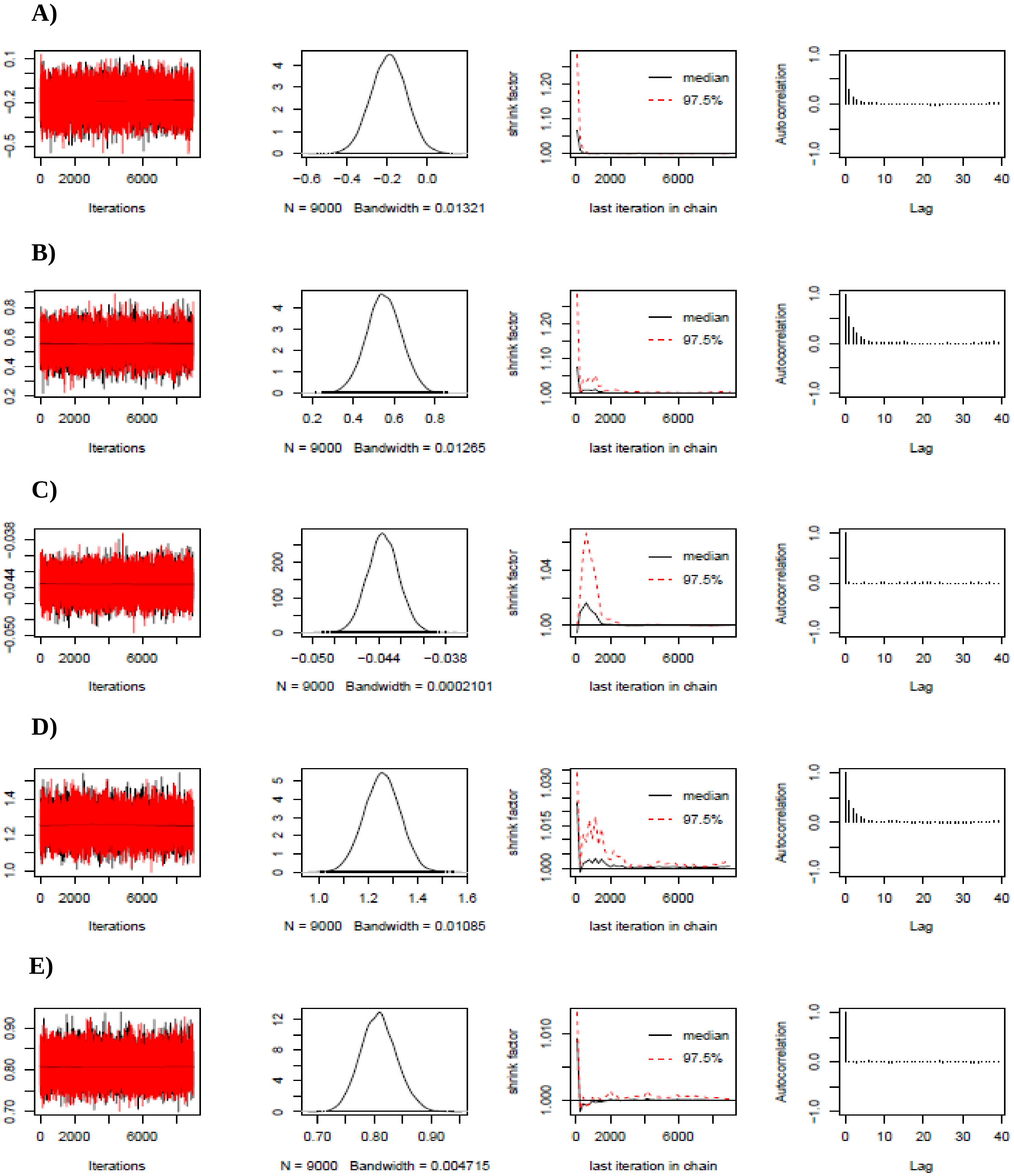


Figure B.6: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for number of specialist visits in model 2 (JA non-cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept

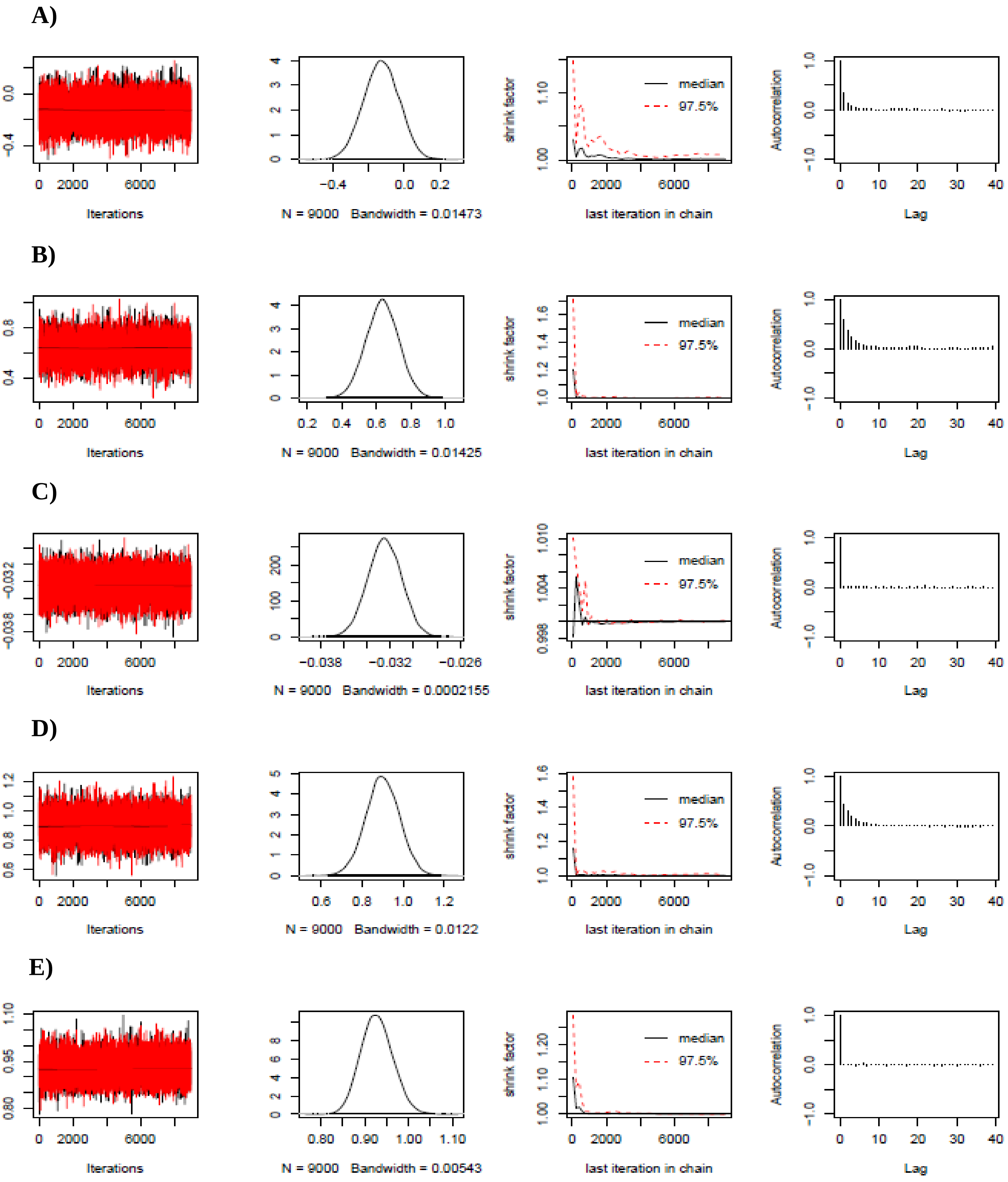


Figure B.7: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for any JA-related visit in model 3 (JA cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept

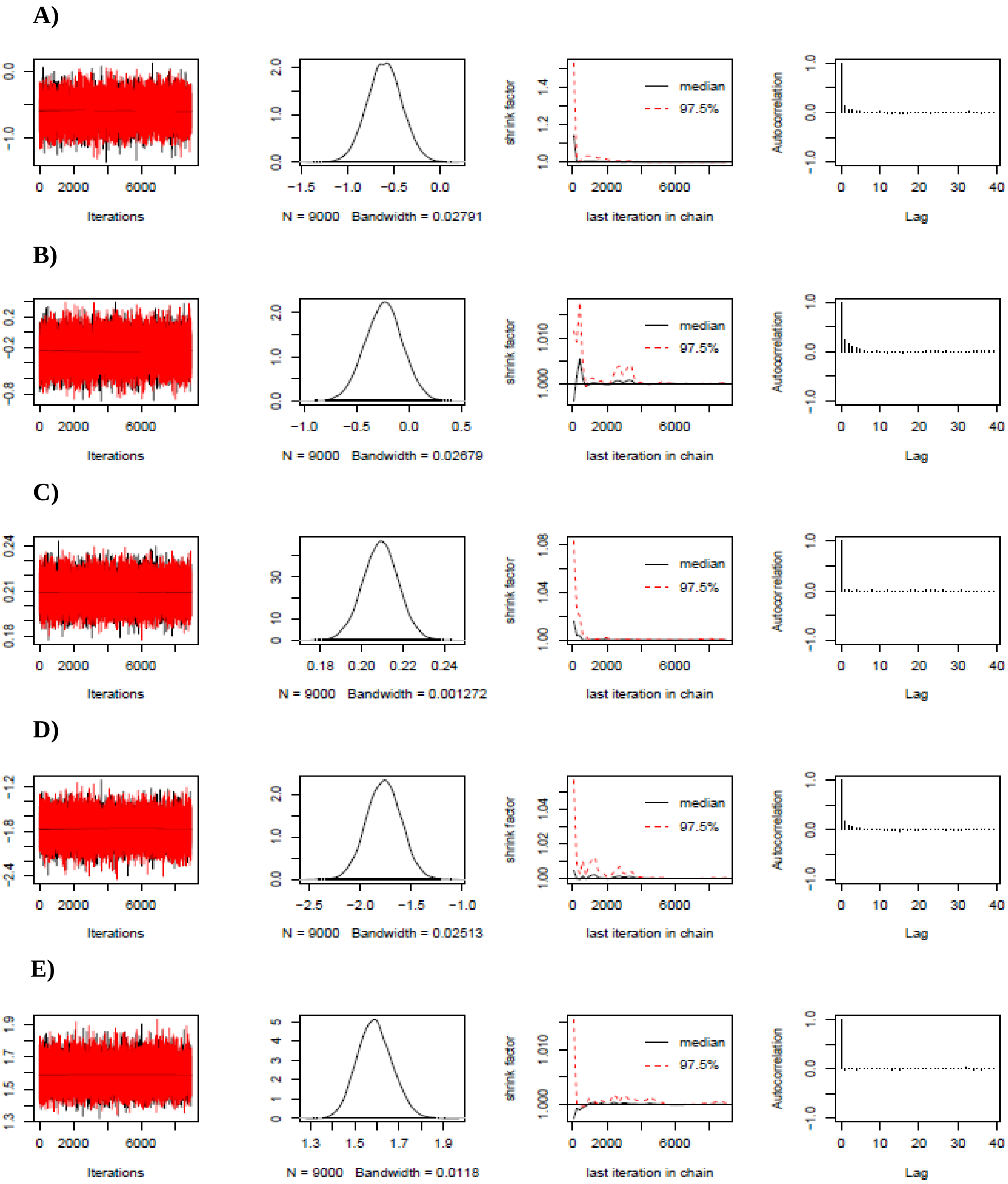


Figure B.8: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for any JA-related visit in model 3 (JA non-cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept

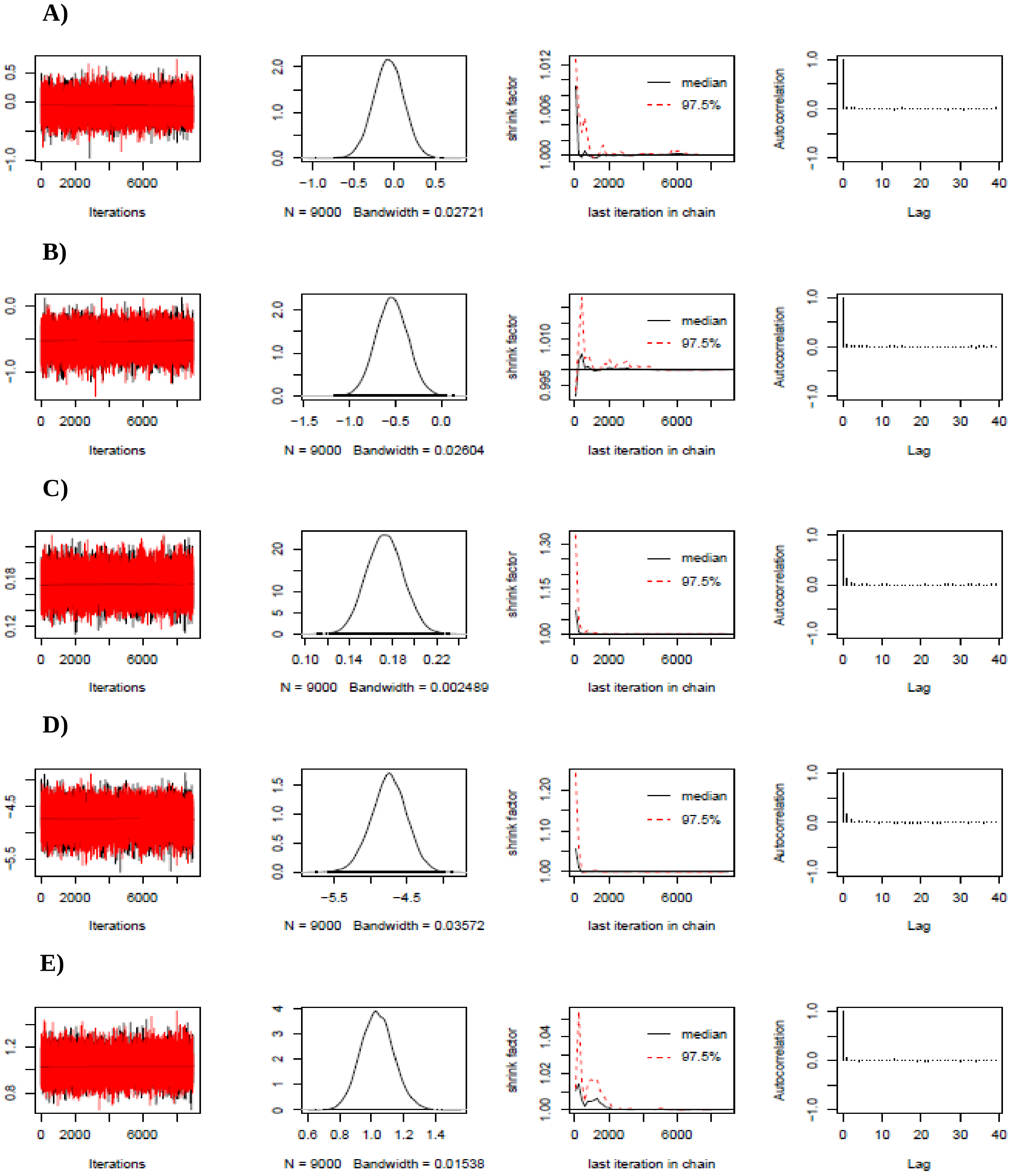


Figure B.9: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for number of specialist visits in model 3 (JA cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept

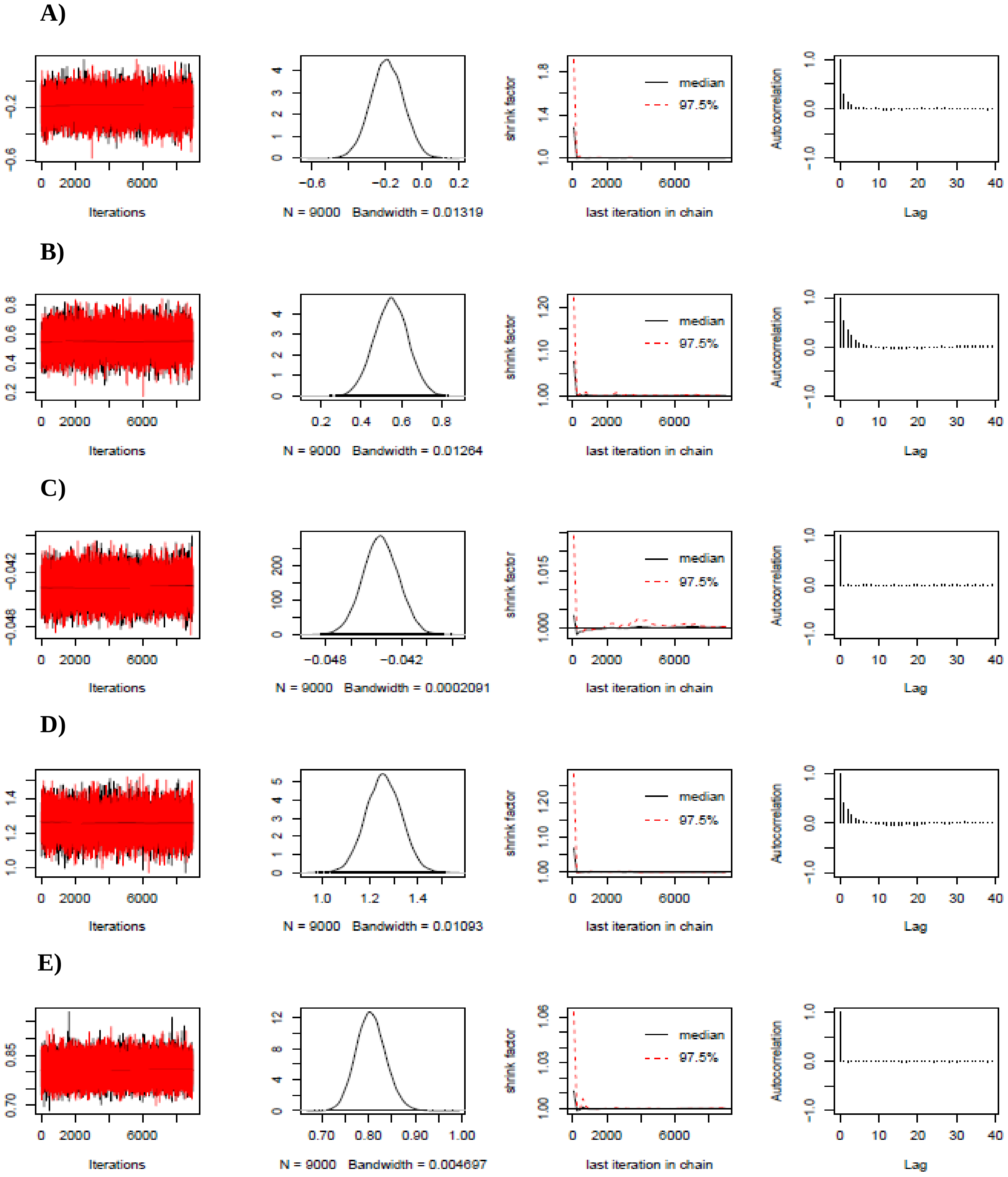


Figure B.10: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for number of specialist visits in model 3 (JA non-cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept

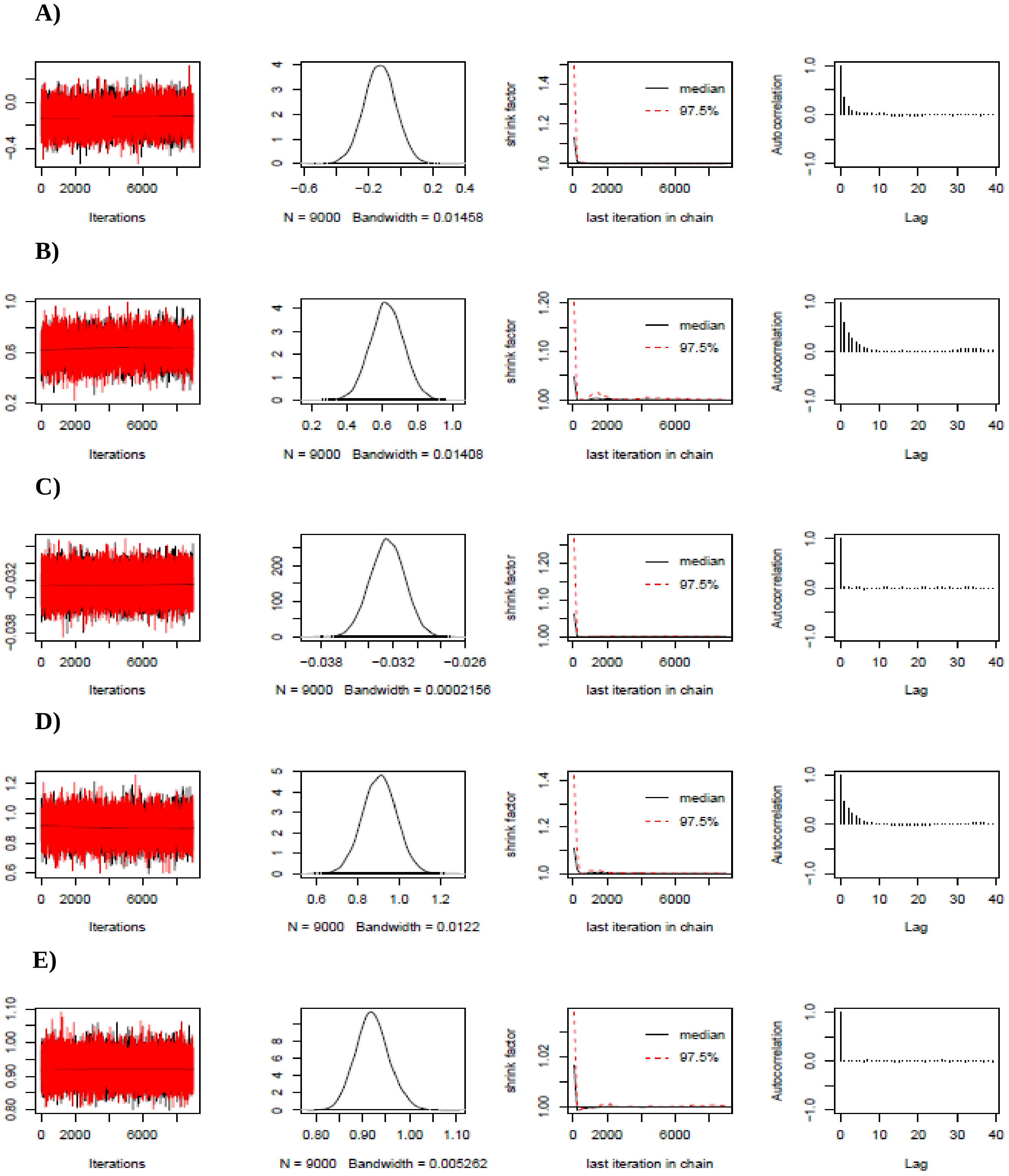


Figure B.11: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for hospitalization in model 3 (JA cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept

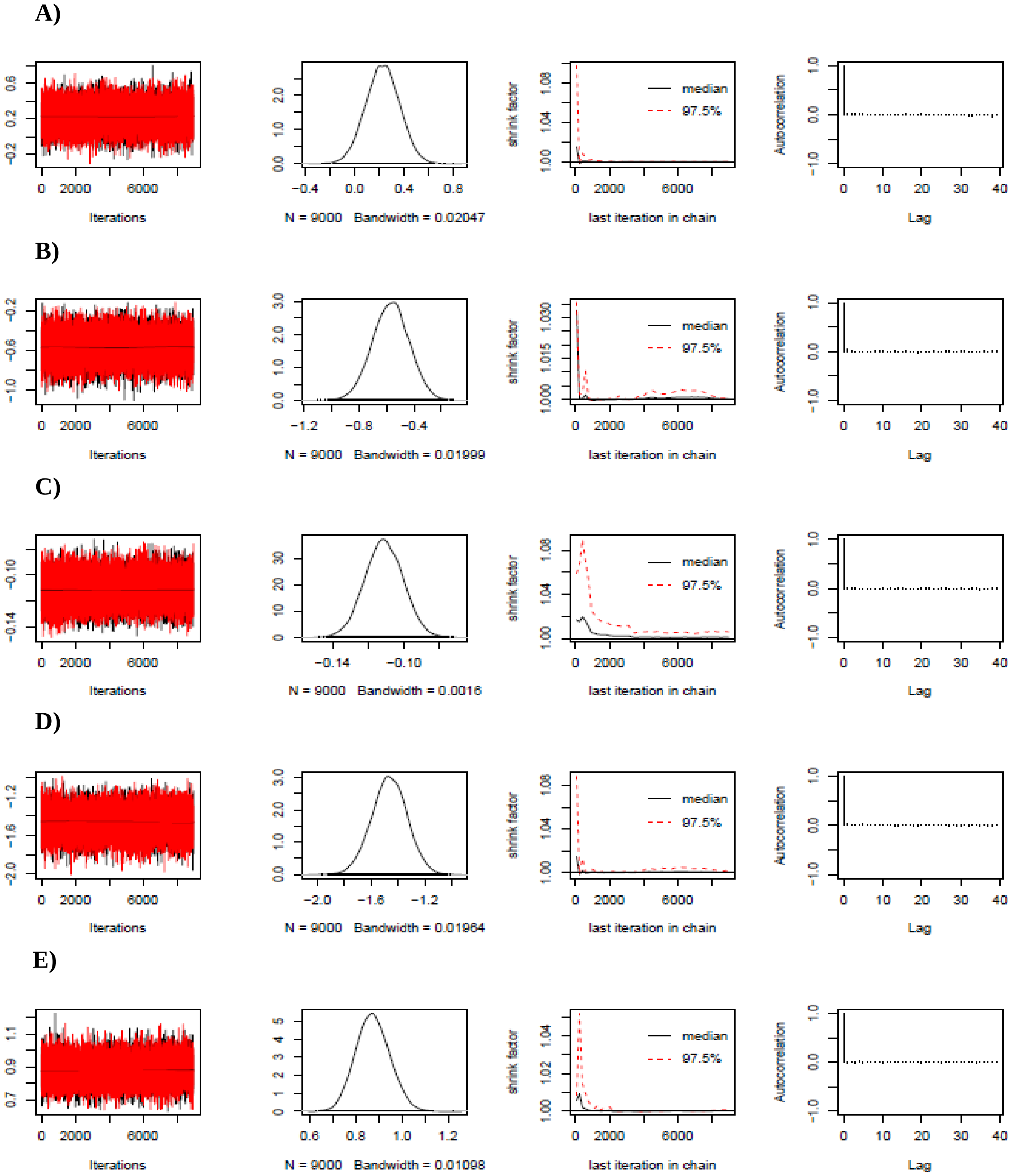
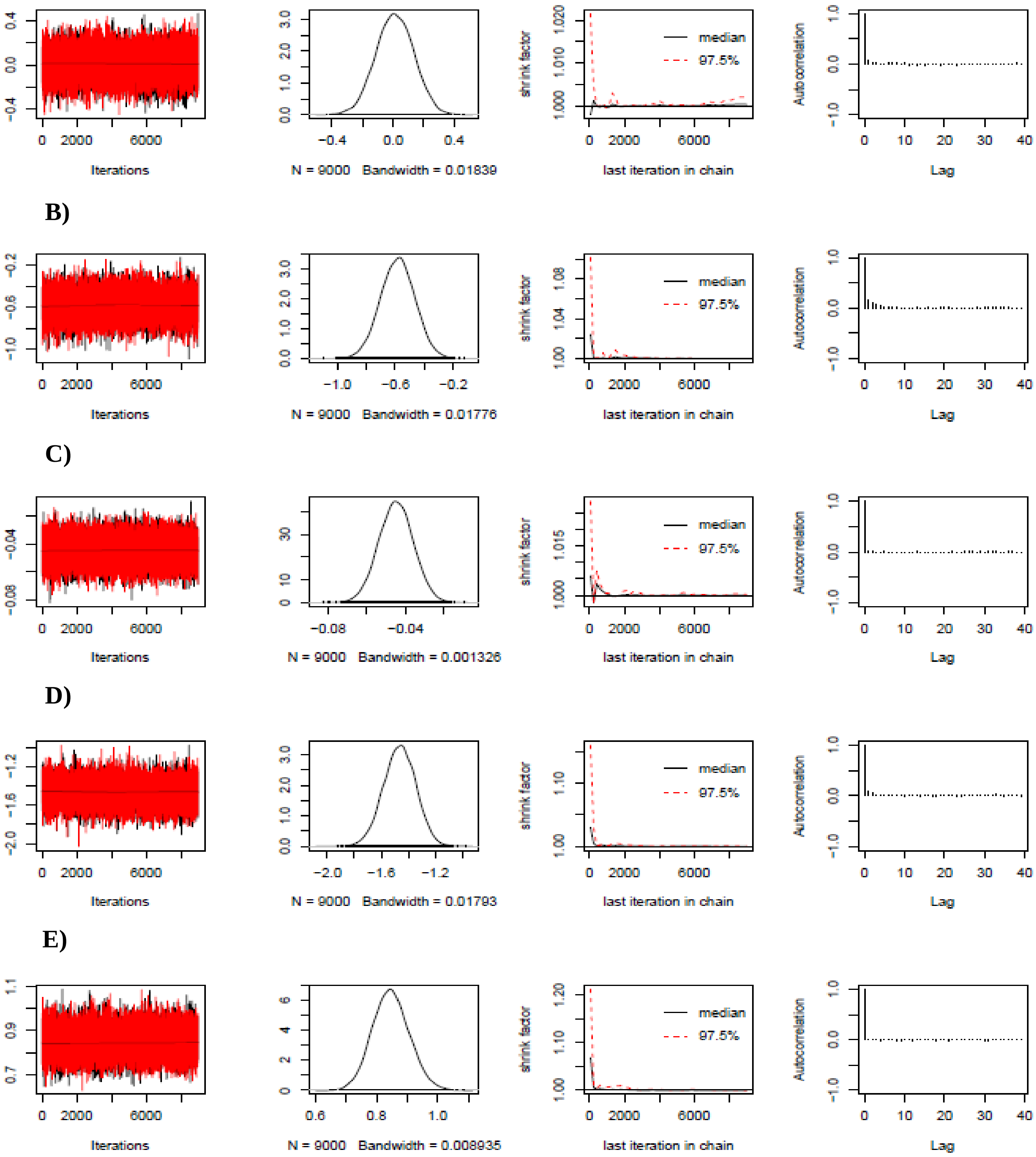


Figure B.12: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for hospitalization in model 3 (JA non-cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept



A)

B)

C)

D)

E)

Figure B.13: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for any JA-related visit in model 4 (JA cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept

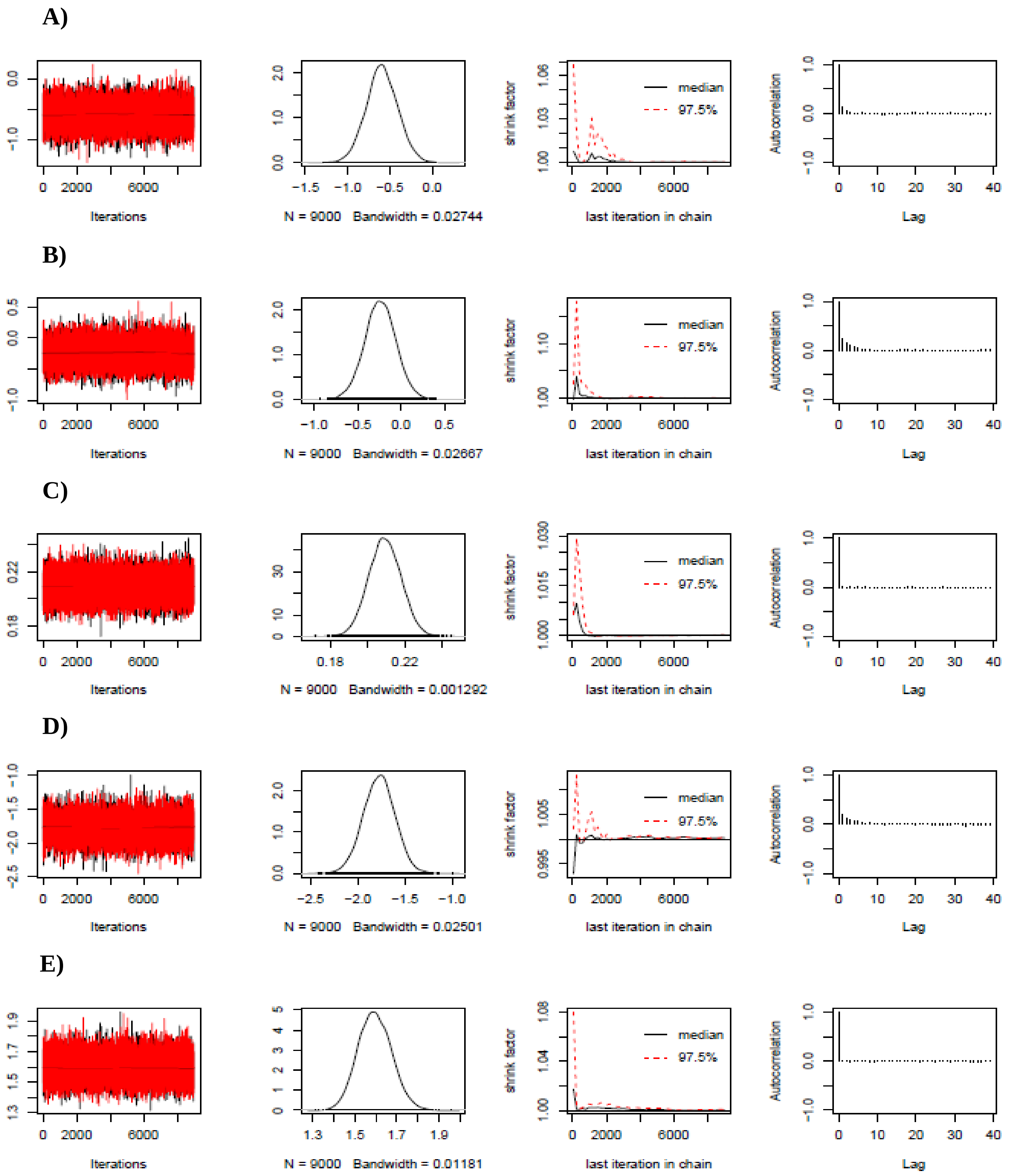


Figure B.14: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for any JA-related visit in model 4 (JA non-cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept

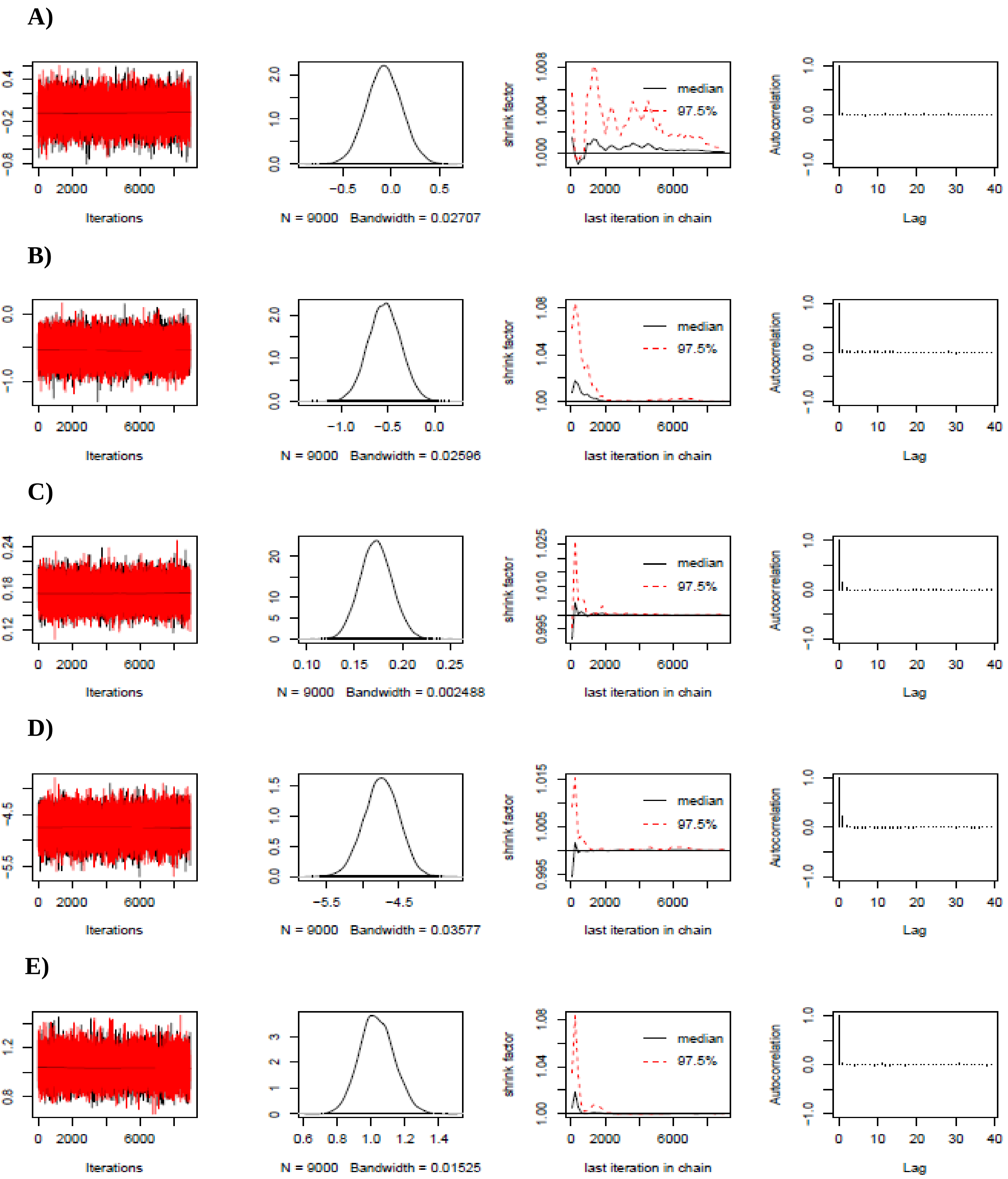


Figure B.15: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for number of specialist visits in model 4 (JA cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept

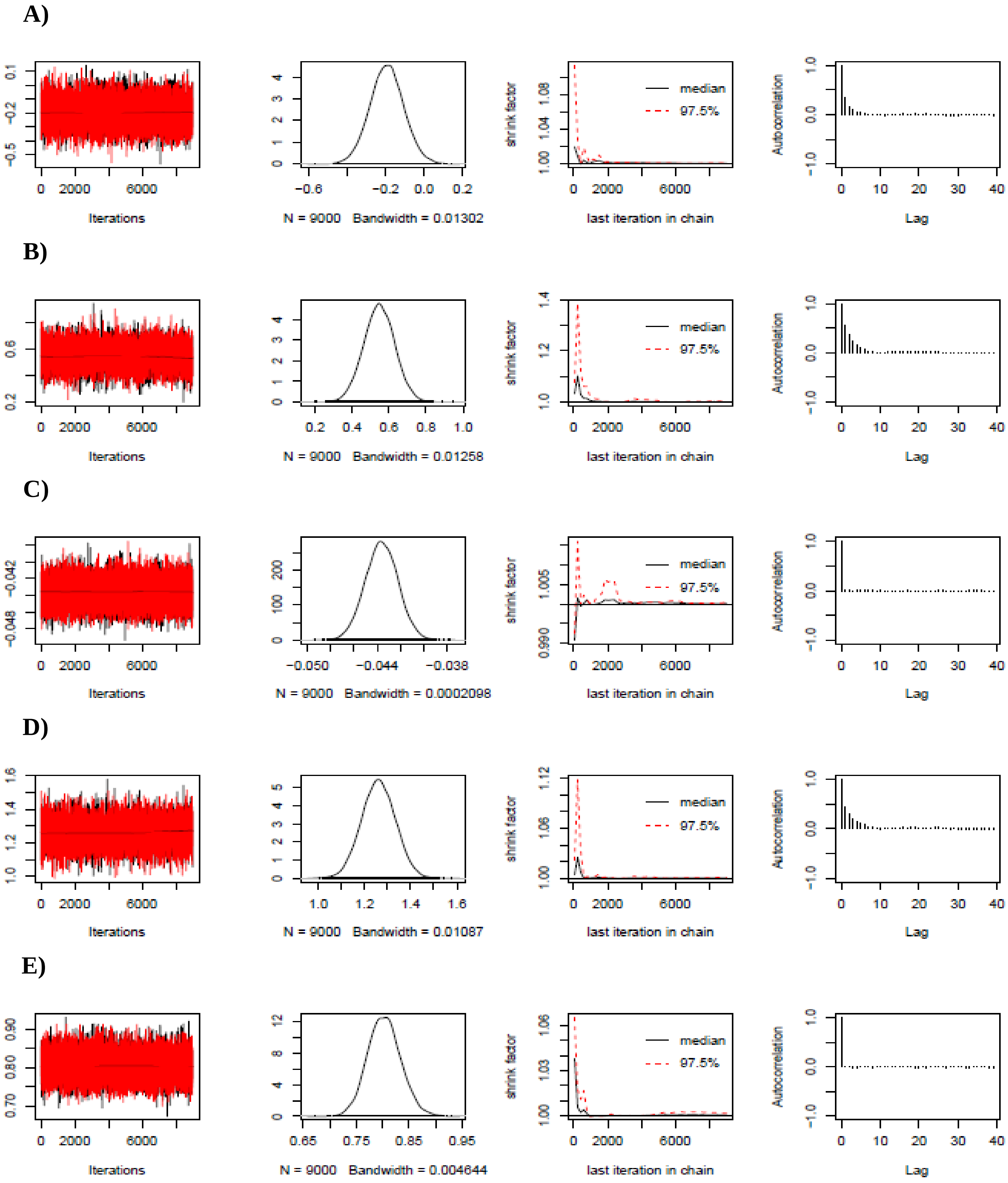


Figure B.16: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for number of specialist visits in model 4 (JA non-cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept

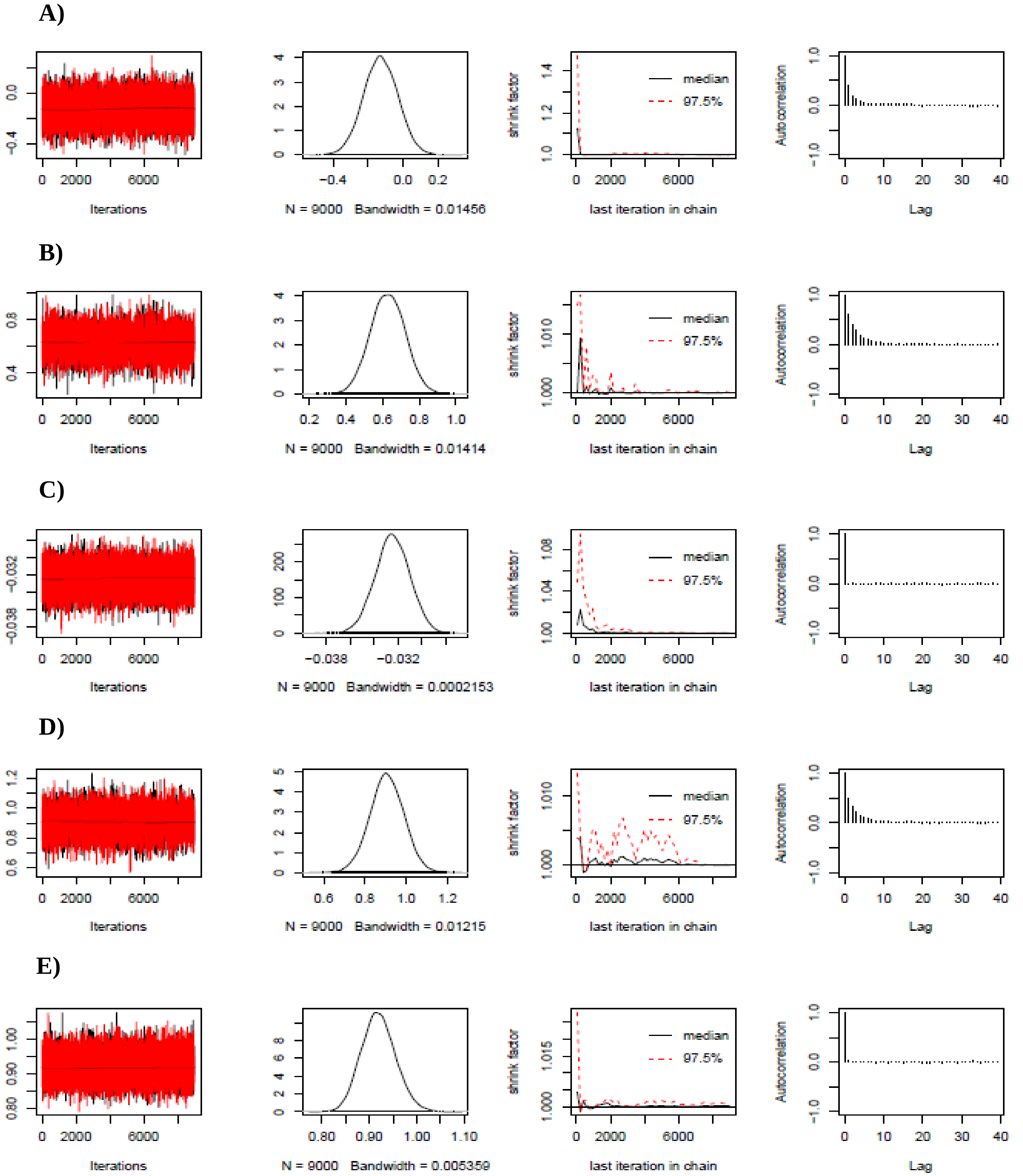


Figure B.17: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for hospitalization in model 4 (JA cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept

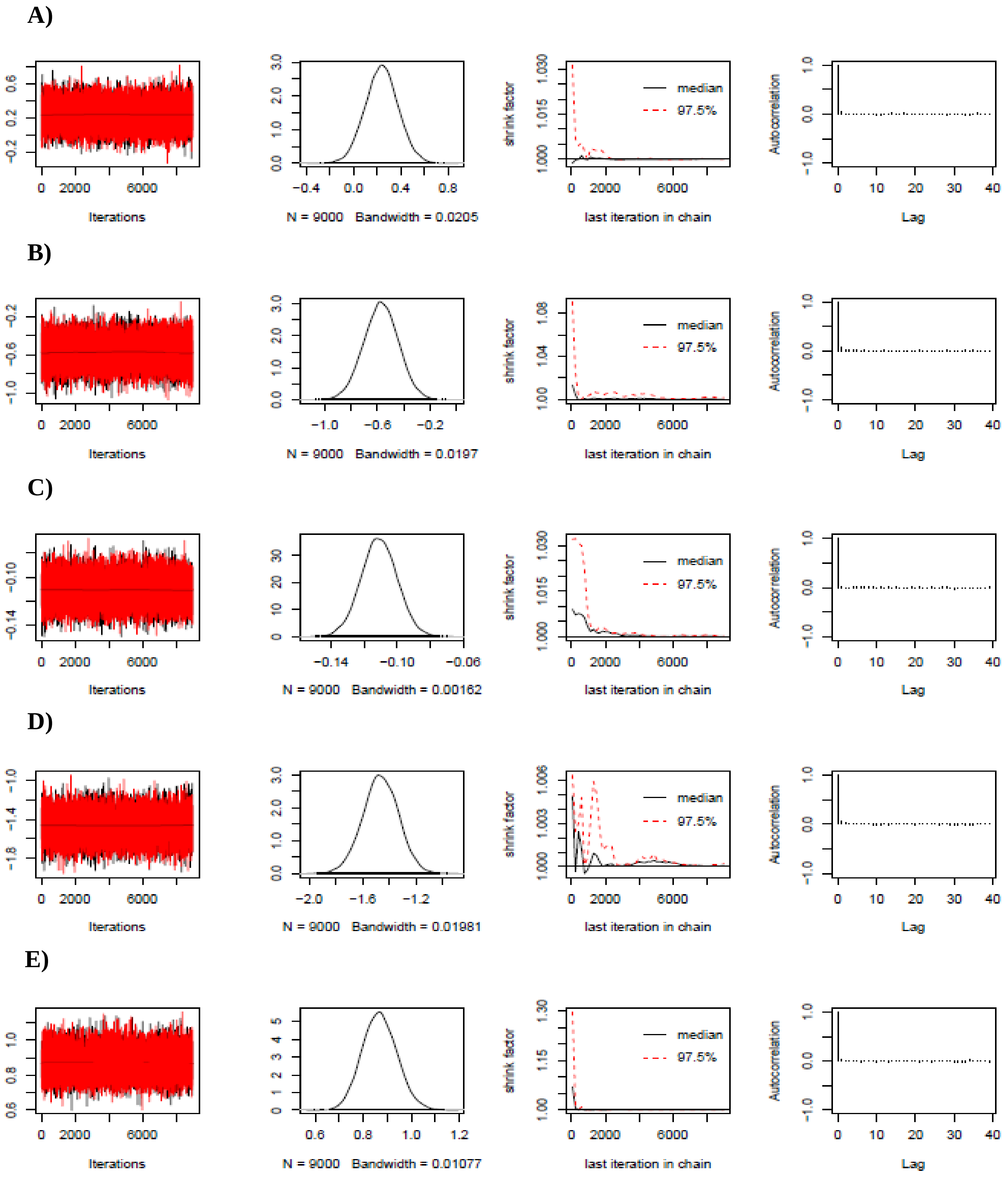


Figure B.18: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for hospitalization in model 4 (JA non-cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept

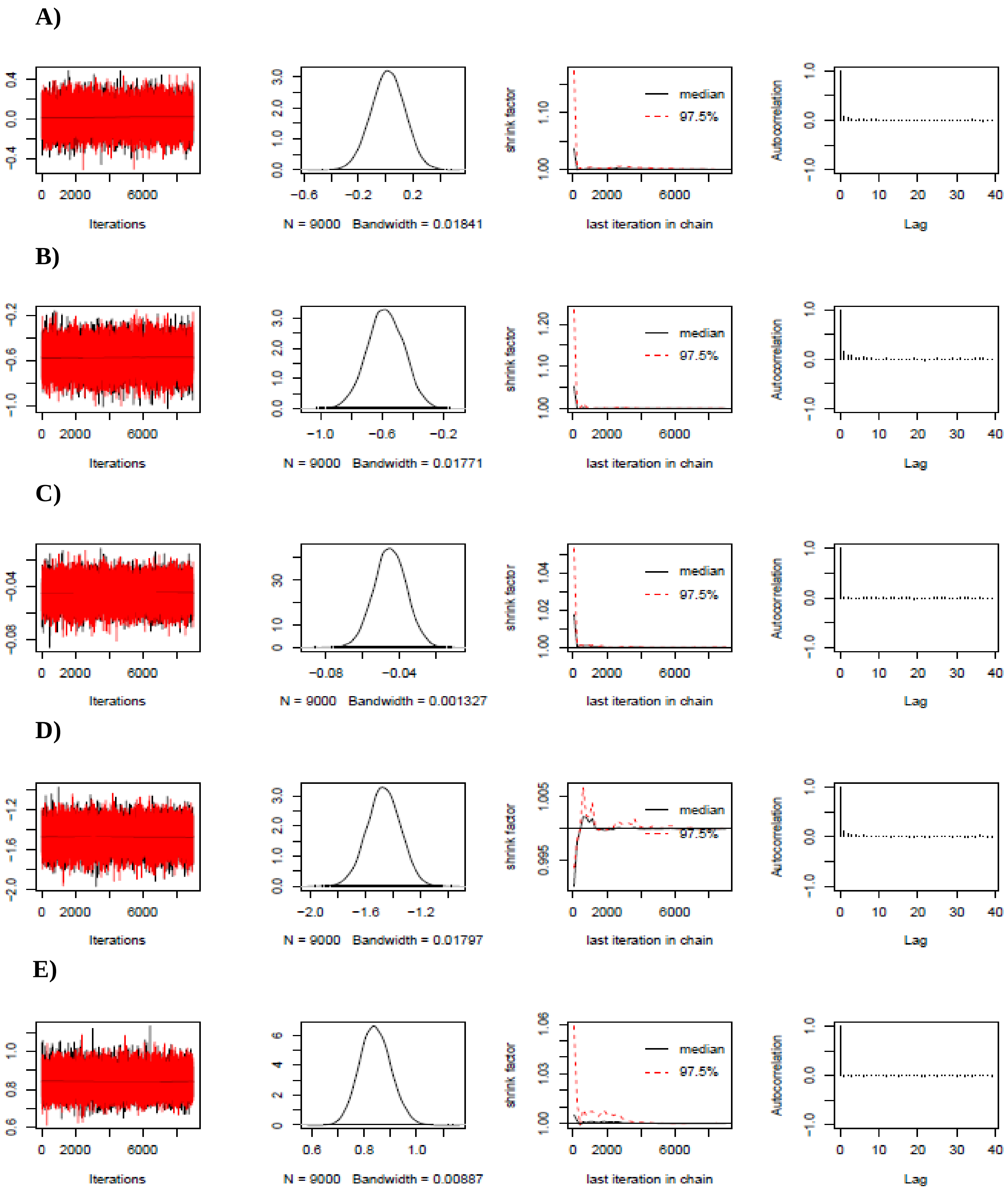
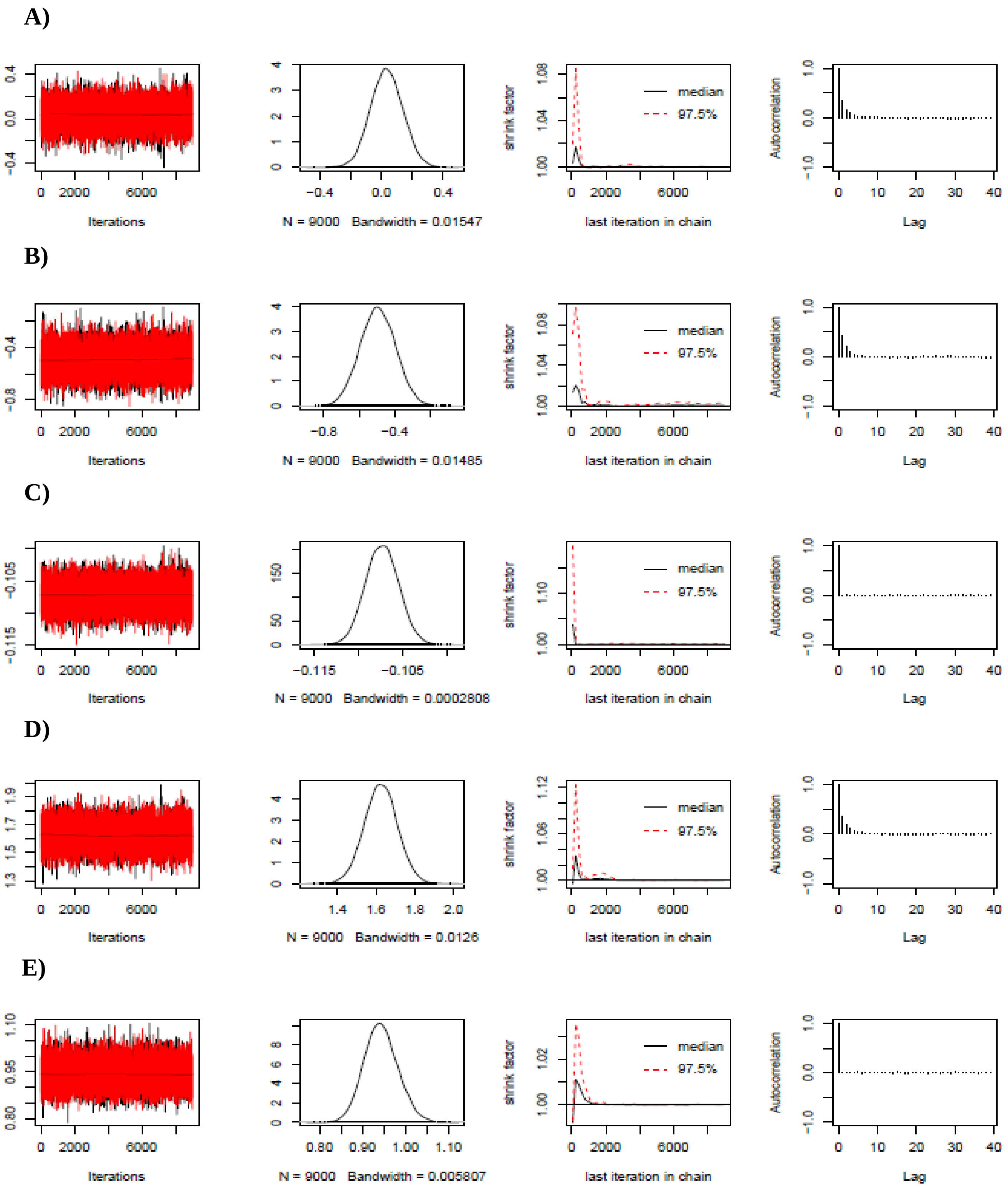


Figure B.19: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for number of general practitioner visits in model 4 (JA cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept


A)

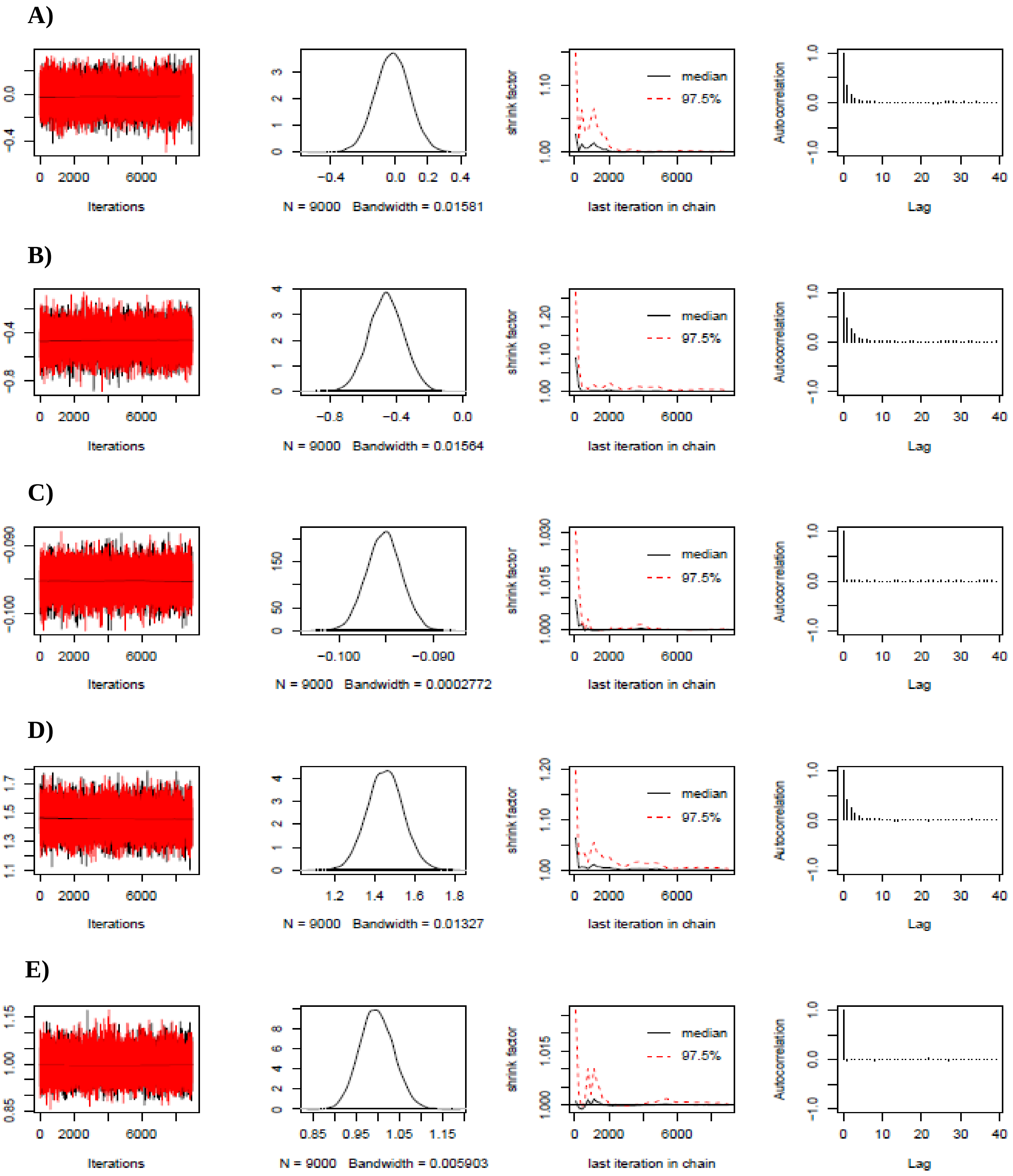
B)

C)

D)

E)

Figure B.20: Trace, density, Gelman-Rubin-Brooks, and autocorrelation plots for the posterior distributions for number of general practitioner visits in model 4 (JA non-cases), A) sex, B) region, C) age, D) expected value of random intercept, E) standard deviation of random intercept


A)

B)

C)

D)

E)

APPENDIX C: CONDITIONAL PREDICTION APPROACH RESULTS

Table C.1: Summary of the classification accuracy for dynamic LoDA using models 1–4 with the conditional prediction approach and naïve prior probabilities

	Model 1[a]	**Model 2**[b]	**Model 3**[c]	**Model 4**[d]
Cut-off	0.47	0.46	0.47	0.48
Sensitivity (classified)*	0.45	0.59	0.69	0.64
Sensitivity	0.45	0.55	0.64	0.61
Specificity (classified)*	0.74	0.69	0.69	0.68
Specificity	0.74	0.65	0.58	0.60
PCC (classified)*	0.61	0.65	0.69	0.66
PCC	0.61	0.60	0.61	0.60
AUC	0.61	0.63	0.63	0.63
PPV	0.66	0.65	0.70	0.67
NPV	0.59	0.64	0.69	0.65
Proportion Unclassified	0.00	0.07	0.11	0.08
Mean Classification Time	2.82	4.53	6.57	5.86

Note: These results were averaged across the five folds

[a] Model 1: Any JA-related visit

[b] Model 2: Any JA-related visit + Number of specialist visits

[c] Model 3: Any JA-related visit + Number of specialist visits + Hospitalization

[d] Model 4: Any JA-related visit + Number of specialist visits + Hospitalization + Number of general practitioner visits

*Measure is calculated without including unclassified individuals in the denominator

PCC: probability of correct classification

AUC: area under the receiver operating characteristic curve

PPV: positive predictive value

NPV: negative predictive value

Table C.2: Summary of the classification accuracy of dynamic LoDA by the prior probability of being a JA case (0.30, 0.40, 0.50, 0.60, 0.70) using model 1 (any JA-related visit) with the conditional prediction approach

	Prior Probability of JA Case				
	0.30	**0.40**	**0.50**	**0.60**	**0.70**
Cut-off	0.28	0.37	0.47	0.57	0.68
Sensitivity (classified)*	0.50	0.48	0.45	0.50	0.47
Sensitivity	0.50	0.48	0.45	0.50	0.47
Specificity (classified)*	0.68	0.69	0.74	0.65	0.68
Specificity	0.68	0.69	0.74	0.65	0.68
PCC (classified)*	0.58	0.59	0.61	0.58	0.58
PCC	0.58	0.59	0.61	0.58	0.58
AUC	0.61	0.61	0.61	0.61	0.60
PPV	0.65	0.59	0.66	0.58	0.59
NPV	0.59	0.58	0.59	0.58	0.58
Proportion Unclassified	0.00	0.00	0.00	0.00	0.00
Mean Classification Time	2.75	2.70	2.82	2.68	2.76

Note: These results are averaged across the five folds

*Measure is calculated without including unclassified individuals in the denominator

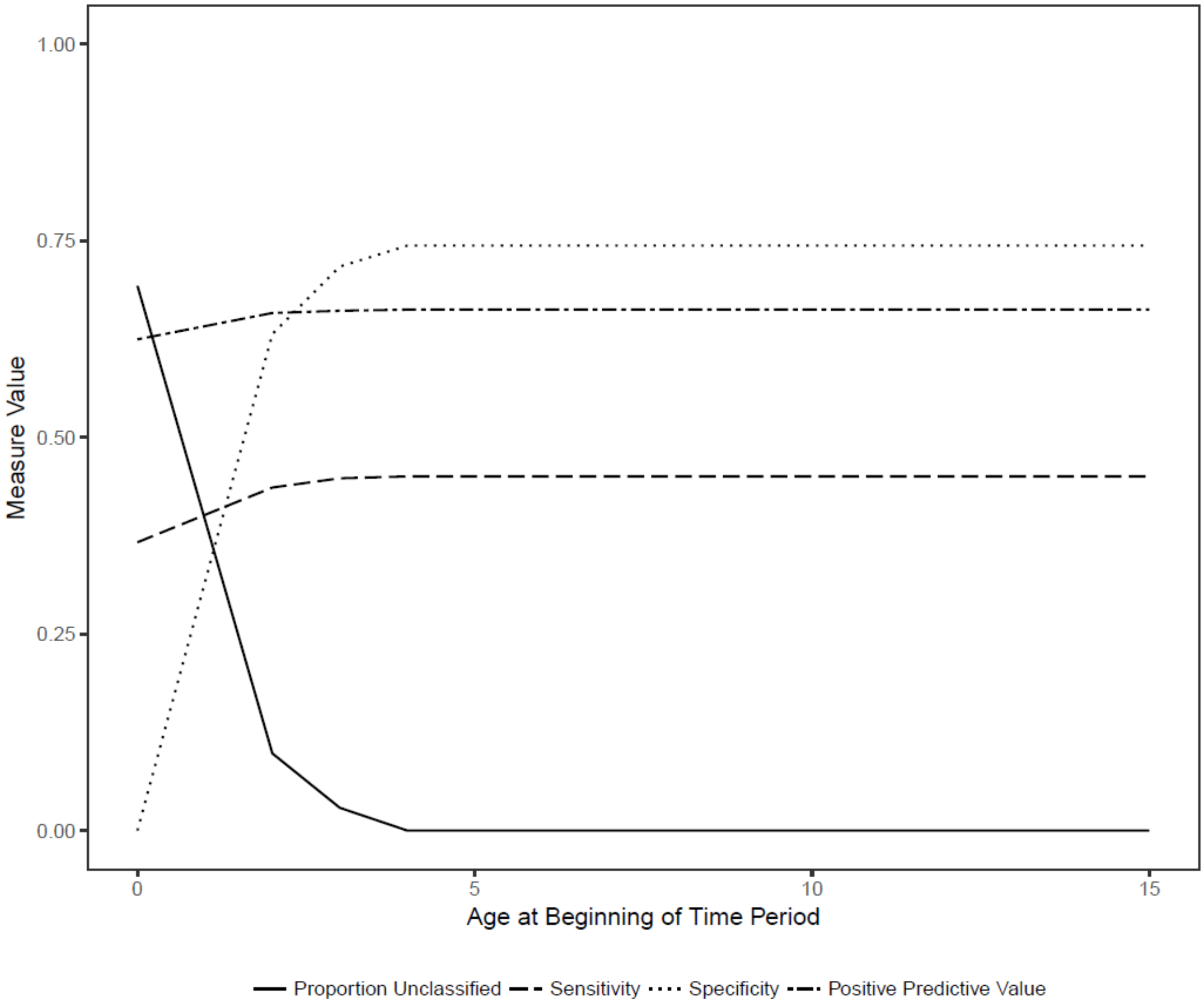
PCC: probability of correct classification

AUC: area under the receiver operating characteristic curve

PPV: positive predictive value

NPV: negative predictive value

Figure C.1: Proportion of unclassified individuals, sensitivity, specificity, and PPV over time periods for dynamic LoDA using model 1 (any JA-related visit) with naïve prior probabilities and the conditional prediction approach



APPENDIX D: DESCRIPTIVE STATISTICS FOR MISCLASSIFED INDIVIDUALS

Table D.1: Characteristics of individuals that were correctly and incorrectly classified by dynamic longitudinal discriminant analysis (LoDA) using model 3 (any JA-related visit) with naïve prior probabilities, n (%)

	Marginal		**Random Effects**		**Conditional**	
	Correct Classification (n = 487)	**Incorrect Classification (n = 310)**	**Correct Classification[a] (n = 624)**	**Incorrect Classification (n = 173)**	**Correct Classification (n = 483)**	**Incorrect Classification (n = 314)**
Sex						
Male	163 (33.5)	116 (37.4)	223 (35.7)	56 (32.4)	159 (32.9)	120 (38.2)
Female	324 (66.5)	194 (62.6)	401 (64.3)	117 (67.6)	324 (67.1)	194 (61.8)
Age at Diagnosis						
0–5	169 (34.7)	79 (25.5)	225 (36.1)	23 (13.3)	168 (34.8)	80 (25.5)
6–10	111 (22.8)	89 (28.7)	169 (27.1)	31 (17.9)	111 (23.0)	89 (28.3)
11–15	207 (42.5)	142 (45.8)	230 (36.9)	119 (68.8)	204 (42.2)	145 (46.2)
Period of Diagnosis						
1983–1992	51 (10.5)	38 (12.3)	*	*	53 (11.0)	36 (11.5)
1993–2002	337 (69.2)	181 (58.4)	*	*	336 (69.6)	182 (58.0)
2003–2012	99 (20.3)	91 (29.4)	*	*	94 (19.5)	96 (30.6)
Region of Residence at Birth						
Winnipeg	278 (57.1)	178 (57.4)	358 (57.4)	98 (56.6)	278 (57.6)	136 (43.3)
Non-Winnipeg	209 (42.9)	132 (42.6)	266 (42.6)	75 (43.4)	205 (42.4)	178 (56.7)
Income Quintile						
Q1–Lowest/Not Found	100 (20.5)	85 (27.4)	131 (21.0)	54 (31.2)	98 (20.3)	87 (27.7)
Q2	92 (18.9)	53 (17.1)	114 (18.3)	31 (17.9)	93 (19.2)	52 (16.6)
Q3	97 (19.9)	50 (16.1)	117 (18.8)	30 (17.3)	97 (20.1)	50 (15.9)
Q4	108 (22.2)	64 (20.)	140 (22.4)	32 (18.5)	108 (22.4)	64 (20.4)
Q5–Highest	90 (18.5)	58 (18.7)	122 (19.6)	26 (15.0)	87 (18.0)	61 (19.4)

[a] Includes individuals that were left unclassified (n=18)

*Not displayed due to low cell count (i.e. 1–5)